

01-06-00

A

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Patent Application of)
)
BEN SHEN, LIANGCHENG DU, CESAR)
SANCHEZ, MEI CHEN and DANIEL J.)
EDWARDS)
)
For: BLEOMYCIN GENE CLUSTER)
COMPONENTS AND THEIR USES)
_____)

San Francisco, California

Patent Application
Assistant Commissioner for Patents
Washington, D.C. 20231

By Express Mail No: **EL160743754US**
Dated: January 5, 2000

PATENT APPLICATION TRANSMITTAL

Sir:

Transmitted herewith for filing is the patent application of inventor(s) Ben Shen, Liangcheng Du, Cesar Sanchez, Mei Chen and Daniel J. Edwards, for "BLEOMYCIN GENE CLUSTER COMPONENTS AND THEIR USES." Enclosed are:

1. 83 pages of the specification, including 73 claims and an abstract.
2. 12 sheets of drawings.
3. 56 pages of Sequence Listing.
4. An oath or declaration of the inventors (unsigned).

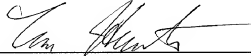
01/05/00
Jc714 U.S. PTO

Jc511 U.S. PTO
09/477962
01/05/00

The filing fee is being deferred at this time.

Dated: January 5, 2000.

Atty. Docket: 2500.125US2
UC Ref: 99-174



Tom Hunter (Reg. No. 38,498)
MAJESTIC, PARSONS, SIEBERT & HSUE P.C.
Four Embarcadero Center, Suite 1100
San Francisco, California 94111-4106
Telephone: (415) 248-5500
Facsimile: (415) 362-5418

In the United States Patent and Trademark Office
U.S. Patent Application For

**BLEOMYCIN GENE CLUSTER COMPONENTS AND THEIR
USES**

Inventor(s): **BEN SHEN**, a citizen of the Peoples Republic of China, residing at
1842 Rushmore Lane, Davis, CA 95616, USA

LIANGCHENG DU, a citizen of Peoples Republic of China, residing
at 1301 Orchard Park Q-9, Davis, CA 95616, USA

CESAR SANCHEZ, a citizen of Spain

MEI CHEN, a citizen of the Peoples Republic of China, residing at
1301 Orchard Park Q-9, Davis, CA 95616, USA

DANIEL J. EDWARDS, a citizen of the United States of America,
residing at 425 Russell Park, Apt. 4, Davis, CA 95616, USA

Assignee: The Regents of the University of California

Entity: Small Entity

MAJESTIC, PARSONS, SIEBERT & HSUE P.C.
Four Embarcadero Center, Suite 1100
San Francisco, CA 94111-4106
Tel: 415 248-5500
Fax: 415 362-5418

BLEOMYCIN GENE CLUSTER COMPONENTS AND THEIR USES

CROSS-REFERENCE TO RELATED APPLICATIONS

This application claims benefit under 35 U.S.C. §119 of provisional applications USSN 60/115,435, filed on January 6, 1999, and USSN 60/118,848, filed on February 5, 1999, both of which are herein incorporated by reference in their entirety for all purposes.

STATEMENT AS TO RIGHTS TO INVENTIONS MADE UNDER FEDERALLY SPONSORED RESEARCH AND DEVELOPMENT

This work was supported in part by an Institutional Research Grant from the American Cancer Society and the School of Medicine, University of California, Davis, National Institutes of Health Grant Number A140475, and a grant from the Searle Scholars Program of the Chicago Community Trust. The Government of the United States of America may have certain rights in this invention.

FIELD OF THE INVENTION

This invention relates to the field of polyketide synthesis and nonribosomal polypeptide synthesis. In particular this invention pertains to the isolation of the bleomycin gene cluster which encodes the first identified hybrid polyketide synthase/nonribosomal peptide synthetase pathway.

BACKGROUND OF THE INVENTION

Polyketides and nonribosomal peptides are two large families of natural products that include many clinically valuable drugs, such as erythromycin and vancomycin (antibacterial), FK506 and cyclosporin (immunosuppressant), and epothilone and bleomycin (BLM) (antitumor). The biosyntheses of polyketides and nonribosomal peptides are catalyzed by polyketide synthases (PKSs) (Hopwood (1997) *Chem. Rev.* 97: 2465; Katz (1997) *Chem. Rev.*, 97: 2557; C. Khosla, (1997) *Chem. Rev.*, 97: 2577; Ikeda and Omura, (1997) *Chem. Rev.*, 97: 2591; Staunton and Wilkinson (1997) *Chem. Rev.*, 97: 2611; Cane *et al.* (1998) *Science* 282: 63) and nonribosomal peptide synthetases (NRPSs) (Cane *et al.* (1998) *Science* 282: 63. Marahiel *et al.* (1997) *Chem. Rev.* 97: 2651; von Döhren *et al.* (1997) *Chem. Rev.* 97: 2675), respectively. Remarkably, PKSs and NRPSs use a very

similar strategy for the assembly of these two distinct classes of natural products by sequential condensation of short carboxylic acids and amino acids, respectively, and utilize the same 4'-phosphopantetheine prosthetic group, via a thioester linkage, to channel the growing polyketide or peptide intermediate during the elongation processes.

Both type I PKSs and NRPSs are multifunctional proteins that are organized into modules. (A module is defined as a set of distinctive domains that encode all the enzyme activities necessary for one cycle of polyketide or peptide chain elongation and associated modifications.) The number and order of modules and the type of domains within a module on each PKS or NRPS protein determine the structural variations of the resulting polyketide and peptide products by dictating the number, order, choice of the carboxylic acid or amino acid to be incorporated, and the modifications associated with a particular cycle of elongation. These features of PKS and NRPS inspired us to search for a hybrid PKS and NRPS system. Since the modular architecture of both PKS (Cane *et al.*(1998) *Science* 282: 63; Katz and Danadio (1993) *Ann. Rev. Microbiol.* 47: 875 (1993); Hutchinson and Fujii (1995) *Ann. Rev. Microbiol.* 49: 201) and NRPS (Cane *et al.*(1998) *Science* 282: 63, Stachelhaus *et al.* (1995) *Science* 269: 69; Stachelhaus *et al.* (198) *Mol. Gen. Genet.* 257: 308; Belshaw *et al.* (1999) *Science* 284, 486) has been exploited successfully in combinatorial biosynthesis of diverse "unnatural" natural products, it is imagined that a hybrid PKS and NRPS system, capable of incorporating both carboxylic acids and amino acids into the final products, could surely lead to even greater chemical structural diversity.

The BLMs, differing structurally at the C-terminal amines of the glycopeptides, are a family of antibiotics produced by *Streptomyces verticillus* (Sv). BLMs exhibit strong antitumor activity through a metal-dependent oxidative cleavage of DNA or RNA in the presence of molecular oxygen and are incorporated into current chemotherapy of several malignancies under the trade name of Blenoxane[®] that contains BLM A2 and BLM B2 as the principal constituents (Sikic *et al.* Eds. (1985) *Bleomycin Chemotherapy*, Academic Press, New York; Natrajan and Hecht (1994) pages 197-242 In: *Molecular Aspects of Anticancer Drug-DNA Interaction Vol. 2*, Neidle and Waring Eds., Macmillan, London). Umezawa, Fujii, Takita, and co-workers extensively studied the biosynthesis of BLM in Sv ATCC15003 by feeding isotope-labeled precursors and by isolating various biosynthetic intermediates and shunt metabolites, establishing that the BLMs are in fact natural hybrid metabolites of polyketide and peptide biosynthesis (Takita and Muroka (1990) pages 289-309 In: *Biochemistry of Peptide Antibiotics: Recent Advances in the Biotechnology of β -Lactams and Microbial Peptides*, Kleinkauf and Von Döhren Eds., W. de

Gruyter, New York). On the assumption that BLM biosynthesis follows the paradigm for peptide and polyketide biosynthesis, we predict that the Blm megasynthetase, which catalyzes the assembly of the BLM backbone from nine amino acids and one acetate, should bear the characteristics of both NRPS and PKS, providing an excellent model to study the mechanism by which NRPS and PKS could be integrated into a productive biosynthetic system to synthesize a hybrid peptide and polyketide metabolite (Fig. 1A) (Shen *et al.* (1999) *Bioorg. Chem.* 27: 155).

SUMMARY OF THE INVENTION

This invention pertains to the isolation and elucidation of the bleomycin gene cluster. Nucleic acid sequences encoding all of the open reading frames (ORFs) that encode polypeptides sufficient to direct the biosynthesis of bleomycin are provided. The nucleic acids can be used in their "native" format or recombined in a wide variety of manners to create novel synthetic pathways.

In one embodiment, this invention provides an isolated nucleic acid comprising a nucleic acid selected from the group consisting of a nucleic acid encoding any one of Blm open reading frames (ORFs) 8 through 41, and/or a nucleic acid encoding a polypeptide encoded by any one of Blm open reading frames (ORFs) 8 through 41, and/or a nucleic acid amplified by polymerase chain reaction (PCR) using any one of the primer pairs identified in Table II and the nucleic acid of a bleomycin-producing organism as a template. The nucleic acid may comprise one or multiple (*e.g.* two, more preferably 3 or more) bleomycin open reading frames (*i.e.* *BLM* ORFs 8 through 41). One preferred nucleic acid comprises a nucleic acid encoding a C domain lacking one or more His residues of the conserved HHxxxDG active site for transpeptidation. In another preferred embodiment the nucleic acid comprises a nucleic acid encoding a protein encoded by a gene selected from the group consisting of blmI, blmII, and blmXI.

In another embodiment this invention provides an isolated nucleic acid encoding a (biosynthetic) module comprising two or more (more preferably three or more, most preferably four or more) catalytic domains of a protein encoded by a nucleic acid of a bleomycin gene cluster wherein said catalytic domains are selected from the group consisting of a condensation (C) domain, an adenylation (A) domain, a peptidyl carrier protein (PCP) domain, a condensation/cyclization domain (Cy), an acyl-carrier protein (ACP)-like domain, an oxidization domain (Ox), a ketoacyl synthase (KS) domain, an acetyl transferase (AT) domain, a ketoreductase (KR) domain, and a methyltransferase (MT) domain. Preferred

nucleic acids comprises a nucleic acid encoding one or more proteins comprising a module selected from the group consisting of NRPS-0, NRPS-1, NRPS-2, NRPS-3, NRPS-4, NRPS-5, NRPS-6, NRPS-7, NRPS-8, NRPS-9, and PKS. Particularly preferred nucleic acids comprise an open reading frame from SEQ ID NO: 1, SEQ ID NO: 2, or SEQ ID NO: 3.

5 In still another embodiment, this invention provides an isolated nucleic acid comprising a nucleic acid encoding a protein encoded by a gene from a BLM gene cluster. Preferred nucleic acids encode a protein encoded by a gene selected from the group consisting of blmI, blmII, and blmXI. In another embodiment, preferred nucleic acids encode a protein encoded by a gene selected from the group consisting of blmIII, blmIV,
10 blmV, blmVI, blmVII, blmIX, and blmX. In still yet another embodiment, the nucleic acid comprises a nucleic acid encoding a protein encoded by blmVIII. Particularly preferred nucleic acids comprise a nucleic acid selected from the group consisting of blmI, blmII, and blmXI. Other particularly preferred nucleic acids comprise a nucleic acid selected from the group consisting of blmIII, blmIV, blmV, blmVI, blmVII, blmIX, and blmX, while still other
15 particularly preferred nucleic acids comprise blmVIII.

In still yet another embodiment, this invention provides an isolated nucleic acid comprising a nucleic acid that encodes a protein comprising at least one catalytic domain selected from the group consisting of a condensation (C) domain, an adenylation (A) domain, a peptidyl carrier protein (PCP) domain, a condensation/cyclization domain (Cy), an
20 acyl-carrier protein (ACP)-like domain, an oxidization domain (Ox), a ketoacyl synthase (KS) domain, an acetyl transferase (AT) domain, a ketoreductase (KR) domain, and a methyltransferase (MT) domain, and that hybridizes to a nucleic acid selected from the group consisting of orf8, orf9, orf10, orf11, orf12, orf13, orf14, orf15, orf16, orf17, orf18, orf19, orf20, orf21, orf22, orf23, orf24, orf25, orf26, orf27, orf28, orf29, orf30, orf31, orf32,
25 orf33, orf34, orf35, orf36, orf37, orf38, orf39, and orf40 under stringent conditions. In certain embodiments this also includes nucleic acids that would stringently hybridizes indicated above, but for, the degeneracy of the nucleic acid code. In other words, if silent mutations could be made in the subject sequence so that it hybridizes to the indicated sequence(s) under stringent conditions, it would be included in certain embodiments. A
30 preferred isolated nucleic acid comprises a nucleic acid encoding a module. A particularly preferred isolated nucleic acid comprises a nucleic acid encoding a BLM gene.

This invention also provides a nucleic acid comprising a nucleic acid selected from the group consisting of consisting of orf8, orf9, orf10, orf11, orf12, orf13, orf14, orf15, orf16, orf17, orf18, orf19, orf20, orf21, orf22, orf23, orf24, orf25, orf26, orf27, orf28,

orf29, orf30, orf31, orf32, orf33, orf34, orf35, orf36, orf37, orf38, orf39, and orf40, or an allelic variant thereof. Preferred nucleic acids comprise a nucleic acid that is a single nucleotide polymorphism (SNP) of a nucleic acid selected from the group consisting of consisting of orf8, orf9, orf10, orf11, orf12, orf13, orf14, orf15, orf16, orf17, orf18,
5 orf19, orf20, orf21, orf22, orf23, orf24, orf25, orf26, orf27, orf28, orf29, orf30, orf31, orf32, orf33, orf34, orf35, orf36, orf37, orf38, orf39, and orf40.

This invention also provides an isolated gene cluster comprising open reading frames encoding polypeptides sufficient to direct the assembly of a bleomycin.

In one embodiment this invention provides an isolated multi-functional
10 protein complex comprising both a polyketide synthase (PKS) and a polypeptide synthetase (NRPS) and/or an isolated nucleic acid encoding a multi-functional protein complex comprising both a polyketide synthase (PKS) and a polypeptide synthetase (NRPS).

This invention also provides various *blm* cluster polypeptides or *blm* cluster-derived polypeptides. Thus, in one embodiment this invention provides an isolated
15 polypeptide comprising a catalytic domain encoded by a nucleic acid of a bleomycin gene cluster wherein said nucleic acid comprises a nucleic acid selected from the group consisting of a nucleic acid encoding any one of Blm open reading frames (ORFs) 8 through 41; and/or a nucleic acid amplified by polymerase chain reaction (PCR) using any one of the primer pairs identified in Table II. Preferred polypeptides comprise an enzymatic domain selected
20 from the group consisting of a condensation (C) domain, an adenylation (A) domain, a peptidyl carrier protein (PCP) domain, a condensation/cyclization domain (Cy), an acyl-carrier protein (ACP)-like domain, an oxidization domain (Ox), a ketoacyl synthase (KS) domain, an acetyl transferase (AT) domain, a ketoreductase (KR) domain, and a methyltransferase (MT) domain. Particularly preferred polypeptides are encoded by the
25 nucleic acids described above and herein.

This invention also provides expression vectors comprising any of the nucleic acids described herein and/or host cells (e.g. *Streptomyces*) transfected and/or transformed with any of these expression vectors. A preferred host cell is transformed with an exogenous nucleic acid comprising a gene cluster encoding polypeptides sufficient to direct the
30 assembly of a bleomycin or bleomycin analog.

This invention also provides methods of use of the *blm* and *blm*-derived nucleic acid(s) and/or polypeptides. One such method is a method of chemically modifying a biological molecule. The method involves contacting a biological molecule that is a substrate for a polypeptide encoded by one or more bleomycin biosynthesis gene cluster

open reading frames with the polypeptide encoded by one or more bleomycin biosynthesis gene cluster open reading frames, whereby the polypeptide chemically modifies the biological molecule. In one particularly preferred embodiment, the biological molecule is an amino acid and said polypeptide is a peptide synthetase. In another preferred embodiment, the polypeptide is a methyl transferase. Other substrates and *blm* encoded polypeptides are illustrated in Table II.

In another embodiment this invention provides a method of coupling a first amino acid to a second amino acid. This method involves contacting the first and second amino acid with a recombinantly expressed bleomycin nonribosomal peptide synthetase (NRPS). A preferred NRPS is selected from the group consisting of NRPS-5, NRPS-4, NRPS-3, NRPS-9, NRPS-8, and NRPS-7. Another preferred NRPS is selected from the group consisting of NRPS-6, NRPS-2, NRPS-1, and NRPS-0. The contacting can be *in vivo* (e.g. in a host cell) or *ex vivo*.

In another embodiment this invention provides a methods of coupling a first fatty acid to a second fatty acid, said method comprising contacting the first and second fatty acids with a recombinantly expressed bleomycin polyketide synthase (PKS). Again, the contacting can be *in vivo* (e.g. in a host cell) or *ex vivo*.

In still another embodiment, this invention provides a method of producing a bleomycin or bleomycin analog. The method involves providing a cell transformed with an exogenous nucleic acid comprising a bleomycin gene cluster encoding polypeptides sufficient to direct the assembly of said bleomycin or bleomycin analog; culturing the cell under conditions permitting the biosynthesis of bleomycin or bleomycin analog; and isolating said bleomycin or bleomycin analog from said cell.

This invention also provides an isolated nucleic acid comprising a nucleic acid encoding a phosphopantetheinyl transferase said nucleic acid encoding a phosphopantetheinyl transferase being selected from the group consisting of: a nucleic acid encoding the protein encoded by the nucleic acid of SEQ ID NO:3; a nucleic acid amplified by polymerase chain reaction (PCR) using primers that specifically amplify ORF 41 (primers: SEQ ID NO:71 and SEQ ID NO:72) and *Streptomyces* nucleic acid as a template; a nucleic acid encoding a polypeptide having phosphopantetheinyl transferase activity where said nucleic acid specifically hybridizes to the nucleic acid of SEQ ID NO: 3 under stringent conditions. In one embodiment, the nucleic acid comprises the nucleic acid of SEQ ID NO:3.

In another embodiment, this invention provides a polypeptide comprising a phosphopantetheinyl transferase encoded by SEQ ID NO:3 or a polypeptide having phosphopantetheinyl transferase activity and the sequence encoded by the nucleic acid of SEQ ID NO: 3 or conservative substitutions of that polypeptide.

5 Also provided are vectors comprising a nucleic acid encoding a phosphopantetheinyl transferase (*e.g.*, as described above) and cells transfected with the vector.

This invention also provides a method of converting an apo carrier protein to a holo carrier protein, said method comprising reacting said apo-carrier protein with a recombinant phosphopantetheinyl transferase encoded by SEQ ID NO:3 and coenzyme A thereby producing a holo-carrier protein.

In certain embodiments, this invention specifically excludes one or more of open reading frames 1 through 41. In particularly preferred embodiments, this invention excludes open reading frames 1 through 7 (Orf 1- Orf 7).

15 **DEFINITIONS**

The "polyketide synthases" (PKSs) refers are multifunctional enzymes, related to fatty acid synthases (FASs). PKSs catalyze the biosynthesis of polyketides through repeated (decarboxylative) Claisen condensations between acylthioesters, usually acetyl, propionyl, malonyl or methylmalonyl. Following each condensation, they typically
20 introduce structural variability into the product by catalyzing all, part, or none of a reductive cycle comprising a ketoreduction, dehydration, and enoylreduction on the β -keto group of the growing polyketide chain. PKSs incorporate enormous structural diversity into their products, in addition to varying the condensation cycle, by controlling the overall chain length, choice of primer and extender units and, particularly in the case of aromatic polyketides, regiospecific cyclizations of the nascent polyketide chain. After the carbon
25 chain has grown to a length characteristic of each specific product, it is typically released from the synthase by thiolysis or acyltransfer. Thus, PKSs consist of families of enzymes which work together to produce a given polyketide. Two general classes of PKSs exist. One class, known as Type I PKSs, is represented by the PKSs for macrolides such as
30 erythromycin. These "complex" or "modular" PKSs include assemblies of several large multifunctional proteins carrying, between them, a set of separate active sites for each step of carbon chain assembly and modification (Cortes *et al.* (1990) *Nature* 348: 176; Donadio *et al.* (1991) *Science* 252: 675; MacNeil *et al.* (1992) *Gene* 115: 119). Structural diversity

occurs in this class from variations in the number and type of active sites in the PKSs. This class of PKSs displays a one-to-one correlation between the number and clustering of active sites in the primary sequence of the PKS and the structure of the polyketide backbone. The second class of PKSs, called Type II PKSs, is represented by the synthases for aromatic compounds. Type II PKSs typically have a single set of iteratively used active sites (Bibb *et al.* (1989) *EMBO J.* 8: 2727; Sherman *et al.* (1989) *EMBO J.* 8: 2717; Fernandez-Moreno, *et al.* (1992) *J. Biol. Chem.* 267:19278).

A "nonribosomal peptide synthase" (NRPS) refers to an enzymatic complex of eucaryotic or procaryotic origin, that is responsible for the synthesis of peptides by a nonribosomal mechanism, often known as thiotemplate synthesis (Kleinkauf and von Doehren (1987) *Ann. Rev. Microbiol.*, 41: 259-289). Such peptides, which can be up to 20 or more amino acids in length, can have a linear, cyclic (cyclosporine, tyrocidine, mycobacilline, surfactin and others) or branched cyclic structure (polymyxin, bacitracin and others) and often contain amino acids not present in proteins or modified amino acids through methylation or epimerization.

A "module" refers to a set of distinctive polypeptide domains that encode all the enzyme activities necessary for one cycle of polyketide or peptide chain elongation and associated modifications.

The terms "isolated" "purified" or "biologically pure" refer to material which is substantially or essentially free from components which normally accompany it as found in its native state. With respect to nucleic acids and/or polypeptides the term can refer to nucleic acids or polypeptides that are no longer flanked by the sequences typically flanking them in nature.

The terms "polypeptide", "peptide" and "protein" are used interchangeably herein to refer to a polymer of amino acid residues. The terms apply to amino acid polymers in which one or more amino acid residue is an artificial chemical analogue of a corresponding naturally occurring amino acid, as well as to naturally occurring amino acid polymers. The term also includes variants on the traditional peptide linkage joining the amino acids making up the polypeptide.

The terms "nucleic acid" or "oligonucleotide" or grammatical equivalents herein refer to at least two nucleotides covalently linked together. A nucleic acid of the present invention is preferably single-stranded or double stranded and will generally contain phosphodiester bonds, although in some cases, as outlined below, nucleic acid analogs are included that may have alternate backbones, comprising, for example, phosphoramidate

(Beaucage *et al.* (1993) *Tetrahedron* 49(10):1925) and references therein; Letsinger (1970) *J. Org. Chem.* 35:3800; Sprinzl *et al.* (1977) *Eur. J. Biochem.* 81: 579; Letsinger *et al.* (1986) *Nucl. Acids Res.* 14: 3487; Sawai *et al.* (1984) *Chem. Lett.* 805, Letsinger *et al.* (1988) *J. Am. Chem. Soc.* 110: 4470; and Pauwels *et al.* (1986) *Chemica Scripta* 26: 1419),

5 phosphorothioate (Mag *et al.* (1991) *Nucleic Acids Res.* 19:1437; and U.S. Patent No. 5,644,048), phosphorodithioate (Briu *et al.* (1989) *J. Am. Chem. Soc.* 111 :2321, O-methylphosphoroamidite linkages (see Eckstein, *Oligonucleotides and Analogues: A Practical Approach*, Oxford University Press), and peptide nucleic acid backbones and linkages (see Egholm (1992) *J. Am. Chem. Soc.* 114:1895; Meier *et al.* (1992) *Chem. Int. Ed.*

10 *Engl.* 31: 1008; Nielsen (1993) *Nature*, 365: 566; Carlsson *et al.* (1996) *Nature* 380: 207). Other analog nucleic acids include those with positive backbones (Denpcy *et al.* (1995) *Proc. Natl. Acad. Sci. USA* 92: 6097; non-ionic backbones (U.S. Patent Nos. 5,386,023, 5,637,684, 5,602,240, 5,216,141 and 4,469,863; Angew. (1991) *Chem. Intl. Ed. English* 30: 423; Letsinger *et al.* (1988) *J. Am. Chem. Soc.* 110:4470; Letsinger *et al.* (1994) *Nucleoside*

15 *& Nucleotide* 13:1597; Chapters 2 and 3, ASC Symposium Series 580, "Carbohydrate Modifications in Antisense Research", Ed. Y.S. Sanghui and P. Dan Cook; Mesmaeker *et al.* (1994), *Bioorganic & Medicinal Chem. Lett.* 4: 395; Jeffs *et al.* (1994) *J. Biomolecular NMR* 34:17; *Tetrahedron Lett.* 37:743 (1996)) and non-ribose backbones, including those described in U.S. Patent Nos. 5,235,033 and 5,034,506, and Chapters 6 and 7, ASC

20 Symposium Series 580, *Carbohydrate Modifications in Antisense Research*, Ed. Y.S. Sanghui and P. Dan Cook. Nucleic acids containing one or more carbocyclic sugars are also included within the definition of nucleic acids (see Jenkins *et al.* (1995), *Chem. Soc. Rev.* pp169-176). Several nucleic acid analogs are described in Rawls, C & E News June 2, 1997 page 35. These modifications of the ribose-phosphate backbone may be done to facilitate the

25 addition of additional moieties such as labels, or to increase the stability and half-life of such molecules in physiological environments.

The term "heterologous" as it relates to nucleic acid sequences such as coding sequences and control sequences, denotes sequences that are not normally associated with a region of a recombinant construct, and/or are not normally associated with a particular cell.

30 Thus, a "heterologous" region of a nucleic acid construct is an identifiable segment of nucleic acid within or attached to another nucleic acid molecule that is not found in association with the other molecule in nature. For example, a heterologous region of a construct could include a coding sequence flanked by sequences not found in association with the coding sequence in nature. Another example of a heterologous coding sequence is a

construct where the coding sequence itself is not found in nature (e.g., synthetic sequences having codons different from the native gene). Similarly, a host cell transformed with a construct which is not normally present in the host cell would be considered heterologous for purposes of this invention.

5 A "coding sequence" or a sequence which "encodes" a particular polypeptide (e.g. a PKS, an NRPS, etc.), is a nucleic acid sequence which is ultimately transcribed and/or translated into that polypeptide *in vitro* and/or *in vivo* when placed under the control of appropriate regulatory sequences. In certain embodiments, the boundaries of the coding sequence are determined by a start codon at the 5' (amino) terminus and a translation stop
10 codon at the 3' (carboxy) terminus. A coding sequence can include, but is not limited to, cDNA from procaryotic or eucaryotic mRNA, genomic DNA sequences from procaryotic or eucaryotic DNA, and even synthetic DNA sequences. In preferred embodiments, a transcription termination sequence will usually be located 3' to the coding sequence.

 Expression "control sequences" refers collectively to promoter sequences,
15 ribosome binding sites, polyadenylation signals, transcription termination sequences, upstream regulatory domains, enhancers, and the like, which collectively provide for the transcription and translation of a coding sequence in a host cell. Not all of these control sequences need always be present in a recombinant vector so long as the desired gene is capable of being transcribed and translated.

20 "Recombination" refers to the reassortment of sections of DNA or RNA sequences between two DNA or RNA molecules. "Homologous recombination" occurs between two DNA molecules which hybridize by virtue of homologous or complementary nucleotide sequences present in each DNA molecule.

 The terms "stringent conditions" or "hybridization under stringent conditions"
25 refers to conditions under which a probe will hybridize preferentially to its target subsequence, and to a lesser extent to, or not at all to, other sequences. "Stringent hybridization" and "stringent hybridization wash conditions" in the context of nucleic acid hybridization experiments such as Southern and northern hybridizations are sequence dependent, and are different under different environmental parameters. An extensive guide
30 to the hybridization of nucleic acids is found in Tijssen (1993) *Laboratory Techniques in Biochemistry and Molecular Biology--Hybridization with Nucleic Acid Probes part I chapter 2 Overview of principles of hybridization and the strategy of nucleic acid probe assays*, Elsevier, New York. Generally, highly stringent hybridization and wash conditions are selected to be about 5°C lower than the thermal melting point (T_m) for the specific sequence

at a defined ionic strength and pH. The T_m is the temperature (under defined ionic strength and pH) at which 50% of the target sequence hybridizes to a perfectly matched probe. Very stringent conditions are selected to be equal to the T_m for a particular probe.

An example of stringent hybridization conditions for hybridization of
5 complementary nucleic acids which have more than 100 complementary residues on a filter
in a Southern or northern blot is 50% formamide with 1 mg of heparin at 42°C, with the
hybridization being carried out overnight. An example of highly stringent wash conditions is
0.15 M NaCl at 72°C for about 15 minutes. An example of stringent wash conditions is a
0.2x SSC wash at 65°C for 15 minutes (*see, Sambrook et al. (1989) Molecular Cloning - A*
10 *Laboratory Manual (2nd ed.)* Vol. 1-3, Cold Spring Harbor Laboratory, Cold Spring Harbor
Press, NY, for a description of SSC buffer). Often, a high stringency wash is preceded by a
low stringency wash to remove background probe signal. An example medium stringency
wash for a duplex of, *e.g.*, more than 100 nucleotides, is 1x SSC at 45°C for 15 minutes. An
example low stringency wash for a duplex of, *e.g.*, more than 100 nucleotides, is 4-6x SSC at
15 40°C for 15 minutes. In general, a signal to noise ratio of 2x (or higher) than that observed
for an unrelated probe in the particular hybridization assay indicates detection of a specific
hybridization. Nucleic acids which do not hybridize to each other under stringent conditions
are still substantially identical if the polypeptides which they encode are substantially
identical. This occurs, *e.g.*, when a copy of a nucleic acid is created using the maximum
20 codon degeneracy permitted by the genetic code.

A "library" or "combinatorial library" of polyketides and/or polypeptides is
intended to mean a collection of polyketides and/or polypeptides (or other molecules)
catalytically produced by a PKS and/or NRPS and/or hybrid PKS/NRPS (or other possible
combination of synthetic elements) gene cluster. The library can be produced by a gene
25 cluster that contains any combination of native, homolog or mutant genes from aromatic,
modular or fungal PKSs and/or NRPSs. The combination of genes can be derived from a
single PKS and/or NRPS gene cluster, *e.g.*, *act*, *fren*, *gra*, *tcm*, *whiE*, *gris*, *ery*, or the like,
and may optionally include genes encoding tailoring enzymes which are capable of
catalyzing the further modification of a polypeptide, polyketide, or other molecule.
30 Alternatively, the combination of genes can be rationally or stochastically derived from an
assortment of NRPS and/or PKS gene clusters. The library of polyketides and/or
polypeptides and/or other molecules thus produced can be tested or screened for biological,
pharmacological or other activity.

By "random assortment" is intended any combination and/or order of genes, homologs or mutants which encode for the various PKS and/or NRPS enzymes, modules, active sites or portions thereof derived from aromatic, modular or fungal PKS and/or NRPS gene clusters.

5 By "genetically engineered host cell" is meant a host cell where the native PKS and/or NRPS gene cluster has been altered or deleted using recombinant DNA techniques or a host cell into which a heterologous PKS and/or NRPS and/or hybrid PKS/NRPS gene cluster has been inserted. Thus, the term would not encompass mutational events occurring in nature. A "host cell" is a cell derived from a procaryotic microorganism
10 or a eucaryotic cell line cultured as a unicellular entity, which can be, or has been, used as a recipient for recombinant vectors bearing the PKS, NRPS, and/or hybrid gene clusters of the invention. The term includes the progeny of the original cell which has been transfected. It is understood that the progeny of a single parental cell may not necessarily be completely identical in morphology or in genomic or total DNA complement to the original parent, due
15 to accidental or deliberate mutation. Progeny of the parental cell which are sufficiently similar to the parent to be characterized by the relevant property, such as the presence of a nucleotide sequence encoding a desired PKS, are included in the definition, and are covered by the above terms.

Expression vectors are defined herein as nucleic acid sequences that are direct
20 the transcription of cloned copies of genes/cDNAs and/or the translation of their mRNAs in an appropriate host. Such vectors can be used to express genes or cDNAs in a variety of hosts such as bacteria, bluegreen algae, plant cells, insect cells and animal cells. Expression vectors include, but are not limited to, cloning vectors, modified cloning vectors, specifically designed plasmids or viruses. Specifically designed vectors allow the shuttling of DNA
25 between hosts, such as bacteria-yeast or bacteria-animal cells. An appropriately constructed expression vector preferably contains: an origin of replication for autonomous replication in a host cell, a selectable marker, optionally one or more restriction enzyme sites, optionally one or more constitutive or inducible promoters. In preferred embodiments, an expression vector is a replicable DNA construct in which a DNA sequence encoding a one or more PKS
30 and/or NRPS domains and/or modules is operably linked to suitable control sequences capable of effecting the expression of the products of these synthase and/or synthetases in a suitable host. Control sequences include a transcriptional promoter, an optional operator sequence to control transcription and sequences which control the termination of transcription and translation, and so forth.

A "bleomycin open reading frame", or "bleomycin ORF", or "*BLM* Orf" refers to a nucleic acid open reading frame that encodes a polypeptide or polypeptide domain that has an enzymatic activity used in the biosynthesis of a bleomycin.

A "PKS/NRPS/PKS" system refers to a synthetic system comprising an NRPS flanked by two PKSs. A "NRPS/PKS/NRPS" system refers to a synthetic system comprising a PKS flanked by two NRPSs. A "hybrid PKS/NRPS system" or a "hybrid NRPS/PKS system" refers to a hybrid synthetic system comprising at least one PKS and one NRPS module. The system can comprise multiple modules and the order can vary.

A "biological molecule that is a substrate for a polypeptide encoded by a bleomycin biosynthesis gene" refers to a molecule that is chemically modified by one or more polypeptides encoded by open reading frame(s) of the *blm* gene cluster. The "substrate" may be a native molecule that typically participates in the biosynthesis of a bleomycin, or can be any other molecule that can be similarly acted upon by the polypeptide.

A "polymorphism" is a variation in the DNA sequence of some members of a species. A polymorphism is thus said to be "allelic," in that, due to the existence of the polymorphism, some members of a species may have the unmutated sequence (*i.e.* the original "allele") whereas other members may have a mutated sequence (*i.e.* the variant or mutant "allele"). In the simplest case, only one mutated sequence may exist, and the polymorphism is said to be diallelic. In the case of diallelic diploid organisms, three genotypes are possible. They can be homozygous for one allele, homozygous for the other allele or heterozygous. In the case of diallelic haploid organisms, they can have one allele or the other, thus only two genotypes are possible. The occurrence of alternative mutations can give rise to triallelic, *etc.* polymorphisms. An allele may be referred to by the nucleotide(s) that comprise the mutation.

"Single nucleotide polymorphism" or "SNPs are defined by their characteristic attributes. A central attribute of such a polymorphism is that it contains a polymorphic site, "X," most preferably occupied by a single nucleotide, which is the site of the polymorphism's variation (Goelet and Knapp U.S. patent application Ser. No. 08/145,145). Methods of identifying SNPs are well known to those of skill in the art (*see, e.g.,* U.S. Patent 5,952,174).

The following abbreviations are used herein:: A, adenylation; ACP, acyl carrier protein; AT, acyltransferase; BLM, bleomycin; C, condensation; Cy, condensation/cyclization; KR, ketoreductase; KS, ketoacyl synthase; MT, methyltransferase; NRPS, nonribosomal peptide synthetase; orf, open reading frame; Ox, oxidation; PCP,

peptidyl carrier protein; PCR, polymerase chain reaction; PKS, polyketide synthase; *Sv*, *Streptomyces verticillus*; ArCP, aryl carrier protein, bp, base pair, CoA, co-enzyme A, DTT, dithiothreitol; FAS, fatty acid synthase; kb, kilobase; PPTase, 4'-phosphopantetheinyl transferase; TCA, trichloroacetic acid; and DEBS, 6-deoxyerythronolide B synthase..

BRIEF DESCRIPTION OF THE DRAWINGS

Figures 1A and 1B illustrate the biosynthetic pathway for bleomycin in *Sv* (ATCC 15003). Figure 1A illustrates a biosynthetic pathway for BLM in *Sv* ATCC15003—intermediates except those in brackets were identified. Figure 1B shows a linear model for the BLM megasynthetase-templated assembly of the BLM peptide/polyketide/peptide aglycone from nine amino acids and one acetate—shaded circles represent atypical domains carrying out the proposed novel chemistry, and arrows with broken line indicate where biosynthetic intermediates were derailed. Three-letter amino acid designations were used. [HO], hydroxylation; [H], reduction.

Figure 2 provides a restriction map and gene organization of the *blm* gene cluster from *Sv* ATCC15003 (B, *Bam*HI). Proposed functions for individual open reading frames are summarized in Tables I and II. Modules for individual NRPS and PKS were given along with their proposed substrates in parentheses.

Figures 3A, 3B, 3C, and 3D illustrate the determination of substrate specificity for NRPS-1 and NRPS-6. Figure 3A shows a comparison of the A3 to A6 region of A domains to 84 NRPS modules available at GenBank that activate various amino acids. Figure 3B shows a comparison of amino acid residues that putatively line the substrate binding pockets for A domains (single-letter amino acid designations were used). The number following the protein name indicates the order of a particular A domain in the multimodular NRPS protein. The protein accession numbers are P48663 (HMWP2), P19828 (AngR), AAC06346 (BacA-2), CAB03756 (MbtB), 3510629 (SyrE-7), 3114612 (AcmB-1), CAA67248 (SnbC-1), and 3560507 (FxbC-2). Dhb stands for 2,3-dehydroaminobutyric acid. It is not known if Dhb is the direct substrate for SyrE-7 or resulted from dehydration of an SyrE-7 activated Thr (Guenzi *et al.* (1998) *J. Biol. Chem.* 273: 32857-32863). Figure 3C illustrates purified proteins after overexpression in *E. coli* as analyzed by electrophoresis on a 10% SDS-polyacrylamide gel (the calculated molecular weights for NRPS-1A and NRPS-6A are 64,212 and 61,899, respectively). Figure 3D illustrates substrate specificities as determined by the ATP-PPi exchange reaction with the amino acids of BLM as substrates

(100% relative activity corresponds to 103,000 cpm for NRPS-1A and 256,000 cpm for NRPS-6A).

Figure 4 illustrates a three-module NRPS/PKS/NRPS model for channeling the growing intermediate between NRPS and PKS modules and between PKS and NRPS modules. The KS, ACP, and C domains are shaded to emphasize their unique activities that are responsible for elongating a growing peptide with a short carboxylic acid and a growing polyketide with an amino acid in hybrid peptide/polyketide/peptide biosynthesis.

Figure 5 illustrates the use of *blmVIII* methyltransferase domain to introduce branched methyl groups in a polyketide synthesis. PCK12 has been described by Kao *et al.* (1995) *J. Am. Chem. Soc.*, 7: 9105-9106. DE-1, DE-2 and DE-3 are three representative products demonstrating the strategy and utility of *blmVIII* in introducing a CH₃ group in polyketide biosynthesis.

Figure 6 illustrates the use of the *blm* NRPS and PKS enzymes to synthesize a variety of hybrid polyketide/peptide molecules including, but not limited to, a family of oxazolines/oxazoles, and thiazoline/thiazoles.

Figure 7 illustrates the use of elements of the *blm* gene cluster to synthesize various sugars.

Figure 8A shows a restriction map of the *blm* gene cluster from *Sv* ATCC15003 (B, *Bam*HI). 8B shows the relative position of the *blmI*, *blmII*, and *blmXI* genes to the two *blmAB* resistance genes (*blm^R*, Bln resistance). Individual open reading frames are represented by open arrows. Figure 8C shows the nucleotide sequence of the *blmI* gene. The potential ribosome-binding site (RBS) and the conserved motif for 4'-phosphopantetheinylation are underlined. The sequence has been deposited into GenBank under accession no. _____.

Figure 9 shows an amino acid sequence comparison of BlnI with PCP domains of known type I NRPSs (Grs-2 [P14688], 36% identity, 58% similarity; SrfA-3 [Q08787], 40% identity, 64% similarity; Vir-s [Y11547], 36% identity, 60% similarity; Sfb [U24657], 40% identity, 54% similarity). Given in brackets are nucleotide sequence accession numbers. The shaded letters indicate similar amino acids. Consensus residues are amino acids that are similar in more than three sequences. The signature motif for 4'-phosphopantetheinylation is underlined.

Figures 10A and 10B show the HPLC analysis of BlnI purified from *E. coli* OG7001(pBS2) (Fig. 10A), and *E. coli* OG7001(pBS2/pDPT-Gsp) (Fig. 10B).

Figure 11 shows the enzyme architecture of type I and type II PKS and NRPS. A, adenylation domain; ACP, acyl carrier protein or ACP domain; AT, acyl transferase; C, condensation protein or C domain; KS, β -ketoacyl synthase domain; KS α , β -ketoacyl synthase α subunit; KS β , β -ketoacyl synthase β subunit; PCP, peptidyl carrier protein or PCP domain.

Figure 12 illustrates the reaction catalyzed by phosphopantetheinyl transferases (PPTases).

Figure 13 shows a restriction map and gene organization of the *pptA* locus from *Sv* ATCC15003

DETAILED DESCRIPTION

Polyketides and polypeptides can be assembled in a remarkably similar manner by repetitive addition of an extending unit to a growing chain by polyketide synthases (PKS) and nonribosomal peptide synthetase (NRPS) respectively. In the case of polyketides, the extending unit is typically a fatty acid (activated as an acyl CoA thioester) while the extending unit for polypeptides is typically an amino acid (activated as an aminonacyl adenylate). Both the PKS and NRPS systems have evolved a modular organization to define the number, sequence, and specificity of the incorporation of the extending unit and utilized the 4'-phosphopantetheine prosthetic group to channel the growing intermediate during the elongation process.

This invention pertains to the discovery that a PKS-bound growing polyketide intermediate could be further elongated by an NRPS module, or conversely, a NRPS-bound growing polypeptide intermediate can be further elongated by a PKS module. This discovery permits the exploitation of NRPS, PKS, and hybrid NRPS/PKS systems to provide a number of novel hybrid peptide/polyketide metabolites from amino acids and short fatty acids.

It was also a discovery of this invention that this hybrid NRPS/PKS/NRPS system is exemplified by the bleomycin (Blm) biosynthesis pathway in *Streptomyces verticillus* (Sv.) (ATCC 15003). The bleomycins are a family of glycopeptide-derived antibiotics originally isolated by Umezawa in 1996 from the fermentation broth of *S. verticillus*. Bleomycins (BLMs) exhibit strong anti-tumor activity are currently used in the treatment of lymphoma, particularly Hodgkin's disease, testicular tumors, squamous cell carcinomas of skin, head, cervix, penis, rectum, and for intracavitary therapy of malignant effusions in ovarian and breast cancer. The commercial product, Blenoxane®, contains

BLM A2 and B2 as the principle constituents. Almost uniquely among anticancer drugs, BLM does not cause myelosuppression, promoting its wide application in combination chemotherapy.

In one aspect, this invention provides a cloned and characterized BLM gene cluster consisting of characteristic NRPS and PKS genes from the Blm producer *Streptoveticillum sp.* (ATCC 15003). The cloned and isolated Blm gene cluster provides a method of recombinantly expressing bleomycin and/or bleomycin analogues. Thus, in one embodiment, this invention provides for nucleic acids encoding bleomycin synthetic machinery or subunits thereof, for cells recombinantly modified to express a bleomycin and/or bleomycin analogue, and for a bleomycin or bleomycin analogue recombinantly expressed in such cells.

Like other polyketide synthase or nonribosomal peptide synthetases, the bleomycin synthetic pathway is organized into modules, each module catalyzing the addition and/or modification of one subunit (e.g. fatty acid or amino acid). Each module is organized into a number of domains each domain having a characteristic activity (e.g. activation, condensation, condensation/cyclization, etc.). The catalytic domains within a module and the modules themselves are often arranged collinearly and the order of biosynthetic modules from NH₂- to COOH-terminus on each PKS and NRPS polypeptide and the number and type of catalytic domains within each determine the order of structural and functional elements in the resulting product. The size and complexity of the ultimately formed product are controlled by the number of repeated acyl chain extension steps that are, in turn, a function of the number and placement of carrier protein domains in these multimodular enzymes. The number composition and order of such domains can be altered either to introduce modifications, e.g. into the bleomycin to produce bleomycin analogues, or to produce different or completely new molecules. Such "recombination" is not restricted solely to recombination among the bleomycin catalytic domains and/or modules, but can also involve recombination between bleomycin modules and/or subunits and other PKS and/or NRPS modules and/or subunit. Moreover the discovery that synthetic pathways can incorporate both PKS and NRPS modules and/or catalytic domains makes available hybrid PKS/NRPS syntheses.

Thus, in one embodiment this invention contemplates the use of *blm* gene cluster modules and/or catalytic domains to make various peptide and/or polyketide, and/or hybrid polypeptide/polyketide metabolites (including, but not limited to bleomycin

intermediates or shunt metabolites), in combinatorial biosynthesis with other polyketide synthases and/or other nonribosomal peptide synthetases.

The *blm* gene cluster contains several glycosylases which can be used alone or in context with other PKS and/or NRPS modules or catalytic domains to make various metabolites with sugars associated with bleomycins (bleomycin sugars).

In addition, the *blm* gene cluster includes a novel methyltransferase domain that can be used to make polyketide metabolites with methyl branch(s).

The *blm* gene cluster also is characterized by the unusual Cy domains as well as the unprecedented Ox domain (see, e.g. BlmIV and BlmIII NRPSs), providing an efficient biosynthesis for a bithiazole structure. The *blm* gene cluster, *blm* modules, or *blm* catalytic domains can be used either individually or collectively (alone or in combinations with other nonribosomal peptide synthetases or polyketide synthases) to make thiazolidine, thiazoline and thiazole, bi-thiazolidine, bithiazoline, and bithiazole-containing microbial metabolites.

Other uses include, but are not limited to the usage of the *blm* gene cluster/modules/catalytic units (either individually or collectively) or the Blm model to make heterocyclic ring-containing microbial metabolites, such as five member S- and N-containing compounds of the thiazolidine, thiazoline and thiazole family or the O- and N-containing compounds of the oxazolidine, oxazoline, and oxazole family or to make sugars, such L-sugars (with the BlmG epimerase), sugars modified by carbamoyl group (with BlmD), and disaccharides.

This invention also includes the discovery of a novel discrete PCP protein (encoded by the *BlmI* gene). Apo-BlmI can be efficiently modified into holo-BlmI either *in vivo* or *in vitro* by PCP-specific 4'-phosphopantetheine transferases (PPTases) such as Gsp and Sfp. Unlike the PCP domains in type I NRPSs, blmI lacks its cognate A domain and can be aminoacylated by Val-A, an A domain from a completely unrelated type I NRPS. BlmI, therefore, represents the first characterized type II PCP, providing the genetic and biochemical evidence to support the existence of a type II NRPS. The latter system is useful, in a manner analogous to the type I NRPS, i.e., modular NRPS, in the combinatorial manipulation of NRPS proteins to generate novel peptides. This invention also includes the discovery and characterization of a novel PPTase (encoded by the *pptA* gene in Figure 13). This PPTase can be used in engineered biosynthesis of polyketides, peptides, hybrid peptide and polyketide metabolites, hybrid polyketide and peptide metabolites, or the combination of both types of metabolites. The PPTase can also be used in converting apo-peptidyl carrier

proteins (both type I and type II) and acyl carrier proteins (both type I and type II) into the holo-proteins.

The Examples provided herein and the accompanying primers permit one of ordinary skill in the art to isolate the *blm* gene cluster of this invention, its constituent ORFs, various modules, or enzymatic domains. The isolated nucleic acid components can be used to express one or more polypeptide components for *in vivo* (e.g. recombinant) synthesis of one or more polypeptides and/or polyketides as indicated above. It will also be appreciated that the *blm* cluster polypeptides can be used for *ex vivo* assembly of various macromolecules.

I. BLM gene cluster and the PPTase gene.

A) The BLM gene cluster.

The nucleic acids comprising the *blm* gene cluster are identified in Tables I and II and listed in the sequence listing provided herein (SEQ ID NOS: 1 and 2, GenBank Accession numbers AT-149091, AT-210249, AF210311). In particular, Table I identifies genes and functions of open reading frames (ORFs) responsible for the biosynthesis of the hybrid peptide/polyketide/peptide backbone and sugar moieties of bleomycin, while Table II identifies a number of ORFs comprising the *blm* gene cluster, identifies the activity of the catalytic domain encoded by the ORF and provides primers for the amplification and isolation of that orf.

As illustrated in Example 1, the *blm* cluster comprises a PKS module, flanked by several NRPS modules along with several sugar biosynthesis genes and genes encoding other biosynthesis enzymes as well as several resistance and regulatory genes (Table 1).

Table I. Determined functions of ORFs in the bleomycin biosynthesis gene cluster

Gene	Amino acids	Sequence Homolog ¹	Proposed function ^{2,3}
<i>orf8</i>	424	YqeR (BAA12461)	Oxidase
<i>blmC</i>	498	RfaE (AA07904.1)	NDP-glucose synthase
<i>blmI</i>	90	GrsB (P14688)	Type II PCP
<i>blmD</i>	545	NodU (Q53515)	Carbamoyl transferase
<i>blmE</i>	390	RfaF (AAD16056)	Glycosyl transferase
<i>orf13</i>	187	MbtH (O05821)	Unknown
<i>blmII</i>	462	Nrp (CAA98937)	NRPS condensation enzyme
<i>orf15</i>	339	SyrP (1890776)	Regulation
<i>blmII</i>	935	HMWP2 (P48633), McbC (P23185)	A PCP Ox

<i>blmIV</i>	2626	HMWP2 (P48633)	C A PCP Cy A PCP Cy
<i>orf18</i>	638	AsnB (2293165)	Asparagine synthetase
<i>blmF</i>	494	RfbC (Q50864)/BlmOrf1 (507319)	Glycosyl transferase/ β -hydroxylase
<i>blmG</i>	325	YtcB (2293288)	Sugar epimerase
<i>blmV</i>	645	McyB (2708278)	PCP C
<i>blmVI</i>	2675	ACoAS (1658531), PksD (S73014) SnbDE (CAA67249)	A ⁴ <u>ACP</u> C A PCP C A
<i>blmVII</i>	1218	SyrE (3510629)	C A PCP
<i>blmVIII</i>	1841	HMWP1 (CAA73127)	<u>KS</u> AT <u>MT</u> KR <u>ACP</u>
<i>blmIX</i>	1066	SafB (1171128)	C A PCP
<i>blmX</i>	2140	TycC (2623773)	C A PCP C A PCP
<i>blmXI</i>	688	SyrE (3510629)	NRPS condensation enzyme
<i>orf28</i>	239	SC9C7.04C (CAA22716)	Unknown
<i>orf29</i>	582	YvdB (CAB08068)	Transmembrane transporter
<i>orf30</i>	113	SmtB (P30340)	Regulation
<i>orf31</i>	117	PhnA (P16680)	Unknown

1. Protein accession numbers are given in parentheses. 2. Underlined domains contain motifs that are clearly different from known NRPS or PKS domains. 3. This A domain lacks the typical NRPS A1, A2, A4, A8, and A9 motifs and more closely resembles acyl CoA synthases. *ORF1* to *ORF7* were reported by Schmidt (1994) *Gene* 151:17-21, who assigned *ORF2* as *blmA* and *ORF4* as *blmB*.

5

Noteworthy are the genes encoding the NRPS and PKS enzymes. The *blmI*, *blmII*, and *blmXI* genes encode NRPSs with an unusual architecture. In contrast to all known NRPSs, which are of modular organization with each module consisting minimally of a condensation (C), an adenylation (A), and a peptidyl carrier protein (PCP) domain, *BlmI*, *BlmII*, and *BlmXI* are discrete proteins homologous to individual domains of type I NRPSs. We have characterized *BlmI* as a type II PCP (Du and Shen (1999) *Chem. Biol.* 6: 507-517). The *BlmII* and *BlmXI* proteins can serve as candidates for type II condensation enzymes.

The *blmIII*, *blmIV*, *blmV*, *blmVI*, *blmVII*, *blmIX*, and *blmX* genes encode modular NRPSs consisting of domains characteristic for known type I NRPSs, such as the A, PCP, C, and condensation/cyclization (Cy) domains, as well as an unprecedented oxidation (Ox) domain. *BlmVI* is unique among all the *Blm* NRPSs identified. Its N-terminal module (NRPS-5) consists of an atypical A domain, which bears a close resemblance to a family of acyl CoA synthases (Fitzmaurice and Kolattukudy (1997) *J. Bacteriol.* 179: 2608-2615; Fitzmaurice and Kolattukudy (1998) *J. Biol. Chem.* 273: 8033-8039), and an acyl carrier protein (ACP)-like domain. Its C-terminal module is truncated and presumably interacts with *BlmV* to constitute the complete NRPS-3 module (Fig. 1B). Also noteworthy are the C domain of NRPS-3 that lacks both His residues of the conserved HHxxxDG (SEQ ID NO: 4) active site for transpeptidation (Stachelhaus *et al.* (1998) *J. Biol. Chem.* 273: 22773-22781)

and the extra C domain at the C-terminus of BlmV. These unusual features associated with BlmVI and BlmV may play roles in the formation of the β -aminoalaninamide and the pyrimidine moieties of BLM, which are unprecedented in peptide biosynthesis.

The *blmVIII* gene encodes a PKS module consisting of domains characteristic for known PKSs, such as ketoacyl synthase (KS), acyltransferase (AT), ketoreductase (KR), and ACP, with malonyl CoA acting as an extending unit according to sequence comparison of the AT domain (Haydock *et al.* (1995) *FEBS Lett.* 374: 246-248) (Fig. 1B).

The identification of an integrated methyltransferase (MT) domain in the middle of *BlmVIII* is unique, representing the first PKS from actinomycetes that contains an internal MT domain.

Table II. *Blm* gene cluster open reading frames (ORFs) and primers for ORF amplification.

Orf #	Position	Activity	Method	Primers Forward Reverse	Seq ID No
orf-8	76183-77457	Oxygen-independent coproporphyrinogen III oxidase	Gapped-blast comparison ¹	F: ATGAGCCACGCCATCGGA R: TCAGCGCGTTCGGGGGC	5 6
orf-9	74690-76186	ADP-heptose synthase (<i>blmC</i>)	Gapped-blast comparison ¹	F: GTGAACACCGACCTGCC R: TCATGGGTGTCTCCCTC	7 8
orf-10	74421-74693	Peptidyl carrier protein (<i>blmI</i>)	Expression and biochemical characterization. ²	F: ATGAGCGCCCGCGGGGC R: TCACGGTCCCGCTCCCC	9 10
orf-11	72787-74424	Carbamyltransferase (<i>blmD</i>)	Gapped-blast comparison ¹	F: ATGAGCGCCGACCCGTC R: TCATGAGCGGGCCGCGT	11 12
orf-12	71618-72790	ADP-heptose:LPS heptosyl transferase (<i>blmE</i>)	Gapped-blast comparison ¹	F: ATGACCAACCCCATGACC R: TCATGGGTACTCTCTGAT	13 14
orf-13	70983-71546	Homolog of mbtH in the synthesis of mycobactin	Gapped-blast comparison ¹	F: ATGACCAACGACCCGCGG R: TCAGTGC CGGACACGCG	15 16
orf-14	69598-70986	Peptide synthetase (condensation, <i>blmII</i>)	Gapped-blast comparison ¹	F: GTGACCGCCCCCGGCACA R: TCATCGTGGCTCCTCGT	17 18
orf-15	68582-69601	Regulatory gene (homolog of <i>syrP</i>)	Gapped-blast comparison ¹	F: GTGAACCGGACGCGCCC R: TCACGCGCTCACTCGTC	19 20
orf-16	65778-68585	Mutated peptide synthetase-oxidase (NRPS-0, <i>blmIII</i>)	Gapped-blast comparison ¹	F: GTGACGAGCGCCCGGCC R: TCACGGGCGCTCCGTGCG	21 22
orf-17	57901-65781	Peptide synthetase (NRPS-2-1, <i>blmIV</i>)	Expression and biochemical characterization. ²	F: ATGCTGCACGGCGCCGCG R: TCACCTCCGGTCCACCTCC	23 24

orf-18	55899-57815	Asparagine synthetase	Gapped-blast comparison ¹	F: GTGAGGCCCGTGTGCGGC R: TCAGCCACCGTTGCCGCG	25 26
orf-19	54418-55902	Homolog of hydroxylase-dehydrogenase (<i>blmF</i>)	Gapped-blast comparison ¹	F: GTGAAGGACCTCGGCCGG R: TCATCCTCCCGGTGCCGCG	27 28
orf-20	53427-54404	Nucleotide-sugar epimerase (<i>blmG</i>)	Gapped-blast comparison ¹	F: GTGACATGGACCGTGGTG R: TCAGGCATCGGCCCTCCC	29 30
orf-21	51493-53430	Peptide synthetase (NRPS-3CT, <i>blmV</i>)	Gapped-blast comparison ¹	F: ATGCGCGGGCATGACGAC R: TCACGGTGTCTCTCCCTC	31 32
orf-22	43263-51290	Peptide synthetase (NRPS-5-4-3, <i>blmVT</i>)	Expression and biochemical characterization. ²	F: ATGAGCCGGCGCGCCGGC R: TCATGCTCGGTATCGGC	33 34
orf-23	39610-43266	Peptide synthetase (NRPS-6, <i>blmVII</i>)	Expression and biochemical characterization. ²	F: GTGACCACGCCCGCATC R: TCATTCCGGGACCGGGCA	35 36
orf-24	34088-39613	Polyketide synthase (<i>blmVIII</i>)	Gapped-blast comparison ¹	F: ATGAGCCATGCCGACGCG R: TCACAGCACCACCTCTTC	37 38
orf-25	30891-34091	Peptide synthetase (NRPS-7, <i>blmIX</i>)	Gapped-blast comparison ¹	F: ATGACCCCGCGCGCGAC R: TCATCGTCCGCGCTCTT	39 40
orf-26	24406-30894	Peptide synthetase (NRPS-9-8, <i>blmX</i>)	Gapped-blast comparison ¹	F: ATGCCTCGGTGTGCCCGA R: TCATTCCGGCGCACCTCC	41 42
orf-27	22127-24193	Peptide synthetase (condensation, <i>blmXI</i>)	Gapped-blast comparison ¹	F: GTGGGTTCCTCGTCGAGG R: TTACACCTCCGTTTCTC	43 44
orf-28	21367-22086	Phosphatidylserine decarboxylase	Gapped-blast comparison ¹	F: ATGGCAGAGACCTGAAC R: TCACGCCACCGGATCTT	45 46
orf-29	19161-20909	Transmembrane transporter	Gapped-blast comparison ¹	F: GTGAGCTCCCTCGCCGTC R: TCATCGTCGGGCACTCGG	47 48
orf-30	18823-19164	Metal dependent regulatory element	Gapped-blast comparison ¹	F: GTGCCGTTCCCGCTGTAT R: TCACCGGGCACTGACCTC	49 50
orf-31	18660-18307	PHNA homolog	Gapped-blast comparison ¹	F: GTGACCAGAGAACCTCCG R: TCAGACCTCTTGACCAC	51 52
orf-32	17736-9211	Peptide synthetase (NRPS-11-10)	Gapped-blast comparison ¹	F: ATGGCCTCAGACGCTTTG R: TCATTGAGACTCCTCCTC	53 54
orf-33	9214-7859	Putative transporter	Gapped-blast comparison ¹	F: ATGATGAAGTCAAGCCCG R: TCAGTGGCTTACAAGGAG	55 56
orf-34	7797-6784	Homolog of clavaminic acid synthase	Gapped-blast comparison ¹	F: ATGACTGACCTGCCGTTG R: TCACACCAGCAGCGAGGT	57 58
orf-35	6773-6021	Thioesterase	Gapped-blast comparison ¹	F: ATGGATTTCCTCCCTCACC R: TCATGCCCTCACTCGGC	59 60
orf-36	6024-4741	Putative transporter	Gapped-blast comparison ¹	F: ATGACCCGCGCGGTGCAC R: TCACTCCTCGGCTTCGGC	61 62
orf-37	4733-3915	Unknown	Gapped-blast comparison ¹	F: GTGTCCAAGAACGCGGCG R: TCATCGGCTCGCCTCGT	63 64
orf-38	3918-2182	Peptide synthetase (NRPS-12)	Gapped-blast comparison ¹	F: ATGACCTCAACCTGCGG R: TCATCGGCACTCCTCT	65 66
orf-39	2185-1199	Regulatory gene (homolog of <i>SyrP</i>)	Gapped-blast comparison ¹	F: GTGACCGGTTCCGTAAAC R: TCATGAGTCCGCGGAGGT	67 68
orf-	1015-1	Peptide synthetase	Gapped-blast	F: ATGACAGAGGTCCGAGGT	69

40			comparison ¹	R: CCCGGCAACCGCCCTCCC	70
orf-41	On a separate sequence	4'-phosphopantetheinyl transferase (<i>pptA</i>)	Expression and biochemical characterization. ²	F: GTGATCGCGCCCTCCTG R: TTACGGGACGGCGGTCCG	71 72

The Blm megasynthetase comprises nine NRPS modules and one PKS module forming a hybrid NRPS/PKS/NRPS metasyntetase (Fig. 1A). Inspection of the blm gene cluster (Fig. 2) showed that the Blm NRPS and PKS modules apparently are not organized according to the “colinearity rule” for BLM biosynthesis (Fig. 1). Detailed functional organization of the megasynthetase and the BLM synthetic pathway is provided in Example I.

B) PPTase

This invention also provides the gene (*pptA*, Fig. 13) encoding phosphopantetheine transferase (PPTase) (GenBank Accession No: AF210311) (*see*, SEQ ID NO: 3). PPTase converts carrier proteins for the growing acyl chain from inactive apo-forms to functional holo-forms by the covalent attachment of the 4'-phosphopantetheine moiety of coenzyme A to a conserved serine residue of the carrier-protein substrate (*see, e.g.*, Fig. 1A).

Using the sequence information provided herein (*e.g.* primer sequences and PPTase sequence) the PPTase nucleic acids can be routinely isolated according to standard methods (*e.g.* PCR amplification). Detailed protocols for the isolation of the PPTase are provided in Example 3.

Other PPTases can be identified using the probes and primers illustrated in Example 3. Briefly, using a primer to the THC motif (5'-C GGC ATG GTC GGC TCC HTN CAN CAY TG -3', SEQ ID NO: 73, where H= C+A, N = A + C + T + G, Y = C + T, K = G + T, R = A + G, W = T + A), and a primer designed around the typical C-terminal PPTase motif (*e.g.*, KEA-1: 5'-T GCA GCA GAA CAG GAG GCK NYC CCA NKG - 3', SEQ ID NO: 74, and KEA-2: 5'- TG GGT CAG CGG GTA CCA NRC YTT RWA - 3', SEQ ID NO: 75), and using *S. verticillus* chromosomal DNA as template, the set of primers THC/KEA-2 a probe can be amplified (about 250 bp), that specifically binds to a PPTase. Libraries of organisms comprising NRPS, PKS, and/or hybrid PKS/NRPS pathways can be probed for the presence of a PPTase sequence. Once hybridizing clones are identified, the PPTase sequence can be isolated according to standard methods well know to those of skill in the art (*see, e.g.*, Example 3).

C) Isolation/preparation of nucleic acids.

In one embodiment, this invention provides nucleic acids for the recombinant expression of a bleomycin. Such nucleic acids include isolated gene cluster(s) comprising open reading frames encoding polypeptides sufficient to direct the assembly of a bleomycin.

5 In other embodiments of this invention, modified bleomycins (e.g. bleomycin analogs), novel polyketides, polypeptides, and combinations thereof (polyketide/polypeptide hybrids) are created by modifying PKSs and/or NRPSs so as to introduce variations into known polymers synthesized by the enzymes. Such variations may be introduced by design, for example to modify a known molecule in a specific way, e.g. by replacing a single
10 monomeric unit within a polymer with another, thereby creating a derivative molecule of predicted structure. Alternatively, variations can be made randomly, for example by making a library of molecular variants of a known polymer by systematically or haphazardly replacing one or more modules or enzymatic domains in a known PKS or NRPS with a collection of alternative modules or domains. Production of alternative/modified PKSs,
15 NRPSs and hybrid systems is described below.

Using the primer and sequence information provided herein, one of ordinary skill in the art can routinely isolate/clone the PKS and/or NRPS modules and/or enzymatic domains described herein. For example, the PCR primers provided in Table II, above, can be used to amplify any of the orfs identified therein. Moreover, using the sequence
20 information for the *blm* gene cluster provided herein, the design of other primers suitable for the amplification of individual ORFs, combinations of ORFs, genes, etc. is routine.

Typically such amplifications will utilize the DNA of an organism containing the requisite genes (e.g. *Streptomyces verticillus*) as a template. Typical amplification conditions include a PCR mixture consisting of 5 ng of *S verticillus* genomic or plasmid
25 DNA as template, 25 pmoles of each primers, 25 μ M dNTP, 5% DMSO, 2 units of *Taq* polymerase, 1 x buffer, with or without 20% glycerol in a final volume of 50 μ L. PCR is carried out (e.g. on a Gene Amp PCR System 2400 (Perkin-Elmer/ABI)) with a cycling scheme as follows: initial denaturing at 94°C for 5 min, 24-36 cycles of 45 sec at 94°C, 1 min at 60°C, 2 min at 72°C, followed by additional 7 min at 72°C. One of skill will
30 appreciate that optimization of such a protocol, e.g. to improve yield, etc. is routine (see, e.g., U.S. Patent No. 4,683,202; Innis (1990) *PCR Protocols A Guide to Methods and Applications* Academic Press Inc. San Diego, CA, etc). In addition, primer may be designed to introduce restriction sites and so facilitate cloning of the amplified sequence into a vector.

Using the information provided herein other approaches to cloning the desired sequences will be apparent to those of skill in the art. For example, the PKS or NRPS modules or enzymatic domains of interest can be obtained from an organism that expresses the same, using recombinant methods, such as by screening cDNA or genomic libraries, derived from cells expressing the gene, or by deriving the gene from a vector known to include the same. The gene can then be isolated and combined with other desired NRPS and/or PKS modules or domains, using standard techniques. If the gene in question is already present in a suitable expression vector, it can be combined *in situ*, with, e.g., other PKS subunits, as desired. The gene of interest can also be produced synthetically, rather than cloned. The nucleotide sequence can be designed with the appropriate codons for the particular amino acid sequence desired. In general, one will select preferred codons for the intended host in which the sequence will be expressed. The complete sequence can be assembled from overlapping oligonucleotides prepared by standard methods and assembled into a complete coding sequence (see, e.g., Edge (1981) *Nature* 292:756; Nambair *et al.* (1984) *Science* 223: 1299; Jay *et al.* (1984) *J. Biol. Chem.* 259:6311). In addition, it is noted that custom gene synthesis is commercially available (see, e.g. Operon Technologies, Alameda, CA).

Examples of such techniques and instructions sufficient to direct persons of skill through many cloning exercises are found in Berger and Kimmel (1989) *Guide to Molecular Cloning Techniques, Methods in Enzymology* 152 Academic Press, Inc., San Diego, CA (Berger); Sambrook *et al.* (1989) *Molecular Cloning - A Laboratory Manual* (2nd ed.) Vol. 1-3, Cold Spring Harbor Laboratory, Cold Spring Harbor Press, NY; Ausubel (1994) *Current Protocols in Molecular Biology*, Current Protocols, a joint venture between Greene Publishing Associates, Inc. and John Wiley & Sons, Inc., U.S. Patent 5,017,478; and European Patent No. 0,246,864.

II. Expression of blm gene clusters, modules, and enzymatic domains.

As indicated above, in one embodiment this invention provides novel NRPS and PKS genes for the efficient recombinant production of both novel and known polyketides, peptides, and polyketide/polypeptide hybrids by expressing them *in vivo*. In other embodiments, such syntheses are carried out *in vitro*. Even *in vitro* syntheses, however, typically utilize recombinantly expressed PKSs, NRPSs, or enzymatic domains thereof. Thus, it is frequently desirable to express protein components of the PKSs or NRPS described above.

Typically expression of the protein components of the pathway and/or of the products of the NRPS/PKS pathway is accomplished by placing the subject PKS or NRPS nucleic acid(s) in an expression vector, and transfecting a cell with the vector such that the cell expresses the desired product(s).

A) Expression vectors

The choice of vector depends on the sequence(s) that are to be expressed.

Any transducible cloning vector can be used as a cloning vector for the nucleic acid constructs of this invention. However, where large clusters are to be expressed, it phagemids, cosmids, PIs, YACs, BACs, PACs, HACs or similar cloning vectors be used for cloning the nucleotide sequences into the host cell. Phagemids, cosmids, and BACs, for example, are advantageous vectors due to the ability to insert and stably propagate therein larger fragments of DNA than in M13 phage and lambda phage, respectively. Phagemids which will find use in this method generally include hybrids between plasmids and filamentous phage cloning vehicles. Cosmids which will find use in this method generally include lambda phage-based vectors into which cos sites have been inserted. Recipient pool cloning vectors can be any suitable plasmid. The cloning vectors into which pools of mutants are inserted may be identical or may be constructed to harbor and express different genetic markers (see, *e.g.*, Sambrook *et al.*, *supra*). The utility of employing such vectors having different marker genes may be exploited to facilitate a determination of successful transduction.

In preferred embodiments of this invention, vectors are used to introduce PKS, NRPS, or NRPS/PKS genes or gene clusters into host (*e.g. Streptomyces*) cells. Numerous vectors for use in particular host cells are well known to those of skill in the art. For example described in Malpartida and Hopwood, (1984) *Nature*, 309:462-464; Kao *et al.*, (1994), *Science*, 265: 509-512; and Hopwood *et al.*, (1987) *Methods Enzymol.*, 153:116-166 all describe vectors for use in various *Streptomyces* hosts.

In a preferred embodiment, *Streptomyces* vectors are used that include sequences that allow their introduction and maintenance in *E. coli*. Such *Streptomyces/E. coli* shuttle vectors have been described (see, for example, Vara *et al.*, (1989) *J. Bacteriol.*, 171:5872-5881; Guilfoile & Hutchinson (1991) *Proc. Natl. Acad. Sci. USA*, 88: 8553-8557.)

The gene sequences, or fragments thereof, which collectively encode a PKS and/or NRPS and/or PKS/NRPS gene cluster, one or more ORFs, one or more modules, or one or more enzymatic domains of this invention, can be inserted into one or more

expression vectors, using methods known to those of skill in the art. Expression vectors will include control sequences operably linked to the desired NRPS and/or PKS coding sequence or fragment thereof. Suitable expression systems for use with the present invention include systems that function in eucaryotic and prokaryotic host cells. However, as explained above, 5 prokaryotic systems are preferred, and in particular, systems compatible with *Streptomyces* spp. are of particular interest. Control elements for use in such systems include promoters, optionally containing operator sequences, and ribosome binding sites. Particularly useful promoters include control sequences derived from PKS and/or NRPS gene clusters, such as one or more *act* promoters. However, other bacterial promoters, such as those derived from 10 sugar metabolizing enzymes, such as galactose, lactose (*lac*) and maltose, will also find use in the present constructs. Additional examples include promoter sequences derived from biosynthetic enzymes such as tryptophan (*trp*), the beta -lactamase (*bla*) promoter system, bacteriophage lambda PL, and T5. In addition, synthetic promoters, such as the tac promoter (U.S. Patent 4,551,433), which do not occur in nature also function in bacterial host cells. In 15 *Streptomyces*, numerous promoters have been described including constitutive promoters, such as *ermE* and *tcmG* (Shen and Hutchinson, (1994) *J. Biol. Chem.* 269: 30726-30733), as well as controllable promoters such as *actI* and *actIII* (Pleper *et al.*, (1995) *Nature*, 378: 263-266; Pieper *et al.*, (1995) *J. Am. Chem. Soc.*, 117: 11373-11374; and Wiesmann *et al.*, (1995) *Chem. & Biol.* 2: 583-589).

20 Other regulatory sequences may also be desirable which allow for regulation of expression of the PKS replacement sequences relative to the growth of the host cell. Regulatory sequences are known to those of skill in the art, and examples include those which cause the expression of a gene to be turned on or off in response to a chemical or physical stimulus, including the presence of a regulatory compound. Other types of 25 regulatory elements may also be present in the vector, for example, enhancer sequences.

Selectable markers can also be included in the recombinant expression vectors. A variety of markers are known which are useful in selecting for transformed cell lines and generally comprise a gene whose expression confers a selectable phenotype on transformed cells when the cells are grown in an appropriate selective medium. Such 30 markers include, for example, genes that confer antibiotic resistance or sensitivity to the plasmid. Alternatively, several polyketides are naturally colored and this characteristic provides a built-in marker for selecting cells successfully transformed by the present constructs.

The various PKS and/or NRPS clusters or subunits of interest can be cloned into one or more recombinant vectors as individual cassettes, with separate control elements, or under the control of, *e.g.*, a single promoter. The PKS and/or NRPS subunits can include flanking restriction sites to allow for the easy deletion and insertion of other PKS subunits so that hybrid PKSs can be generated. The design of such unique restriction sites is known to those of skill in the art and can be accomplished using the techniques described above, such as site-directed mutagenesis and PCR.

Methods of cloning and expressing large nucleic acids such as gene clusters, including PKS- or NRPS-encoding gene clusters, in cells including *Streptomyces* are well known to those of skill in the art (*see, e.g.*, Stutzman-Engwall and Hutchinson (1989) *Proc. Natl. Acad. Sci. USA*, 86: 3135-3139; Motamedi and Hutchinson (1987) *Proc. Natl. Acad. Sci. USA*, 84: 4445-4449; Grim *et al.* (1994) *Gene*, 151: 1-10; Kao *et al.* (1994) *Science*, 265: 509-512; and Hopwood *et al.* (1987) *Meth. Enzymol.*, 153: 116-166). In some examples, nucleic acid sequences of well over 100kb have been introduced into cells, including prokaryotic cells, using vector-based methods (*see, for example*, Osoegawa *et al.*, (1998) *Genomics*, 52: 1-8; Woon *et al.*, (1998) *Genomics*, 50: 306-316; Huang *et al.*, (1996) *Nucl. Acids Res.*, 24: 4202-4209). In addition, the cloning and overexpression of NRPS-1 and NRPS-6 is illustrated in Example 1.

In certain embodiments this invention may make use of genetically engineered cells that either lack PKS and/or NRPS genes or have their naturally occurring PKS and/or NRPS genes substantially deleted. These host cells can be transformed with recombinant vectors, encoding a variety of PKS and/or NRPS gene clusters, for the production of active polyketides. The invention provides for the production of significant quantities of product, *e.g.* a bleomycin, at an appropriate stage of the growth cycle. The BLMs or other hybrid polyketide/peptide metabolites so produced can be used as therapeutic agents, to treat a number of disorders, depending on the type of metabolites in question. For example, several of the polyketides and peptides produced by the present method will find use as immunosuppressants, as anti-tumor agents, as well as for the treatment of viral, bacterial and parasitic infections. The ability to recombinantly produce polyketides and peptides also provides a powerful tool for characterizing PKSs and/or NRPSs and the mechanism of their actions.

B) Host cells.

The vectors described above can be used to express various protein components of the polyketide and/or polypeptide synthetic modules for subsequent isolation and/or to provide a biological synthesis of one or more desired biomolecules (e.g.

5 polyketides, peptides, *etc.*). Where one or more proteins of the *blm* cluster are expressed (e.g. overexpressed) for subsequent isolation and/or characterization, the proteins are expressed in any prokaryotic or eukaryotic cell suitable for protein expression. In one preferred embodiment, the proteins are expressed in *E. coli*. Overexpression of *blmI* in *E. coli* is described in Example 2.

10 Host cells for the recombinant production of the subject polyketides can be derived from any organism with the capability of harboring a recombinant PKS, NRPS or PKS/NRPS gene cluster. Thus, the host cells of the present invention can be derived from either prokaryotic or eucaryotic organisms. However, preferred host cells are those constructed from the actinomycetes, a class of mycelial bacteria which are abundant
15 producers of a number of polyketides and peptides. A particularly preferred genus for use with the present system is Streptomyces. Thus, for example, *S. verticillus*, *S. ambofaciens*, *S. avermitilis*, *S. azureus*, *S. cinnamonensis*, *S. coelicolor*, *S. curacoi*, *S. erythraeus*, *S. fradiae*, *S. galilaeus*, *S. glaucescens*, *S. hygrosopicus*, *S. lividans*, *S. parvulus*, *S. peucetius*, *S. rimosus*, *S. roseofulvus*, *S. thermotolerans*, *S. violaceoruber*, among others, will provide
20 convenient host cells for the subject invention, with *S. coelicolor* being preferred (*see, e.g., Hopwood, D. A. and Sherman, D. H. Ann. Rev. Genet. (1990) 24:37-66; O'Hagan, D. The Polyketide Metabolites (Ellis Horwood Limited, 1991), for a description of various polyketide-producing organisms and their natural products.*)

In a preferred embodiment, the above-described cells are genetically
25 engineered by deleting one or more naturally occurring PKS and/or NRPS genes therefrom, using standard techniques, such as by homologous recombination. (*see, e.g., Khosla, et al. (1992) Molec. Microbiol. 6: 3237.*)

In certain embodiments, a eukaryotic host cell is preferred (e.g. where certain glycosylation patterns are desired). Suitable eukaryotic host cells are well known to those of
30 skill in the art. Such eukaryotic cells include, but are not limited to yeast cells, insect cells, plant cells, fungal cells, and various mammalian cells (e.g. COS, CHO HeLa cells lines and various myeloma cell lines)

C) Protein/polyketide recovery.

Polypeptide and/or polyketide recovery is accomplished according to standard methods well known to those of skill in the art. Thus, for example where *blm* cluster proteins are to be expressed and isolated, the proteins can be expressed with a convenient tag to facilitate isolation (e.g. a His₆) tag. Other standard protein purification techniques are suitable and well known to those of skill in the art (see, e.g., Quadri *et al.* (1998) *Biochemistry* 37: 1585-1595; Nakano *et al.* (1992) *Mol. Gen. Genet.* 232: 313-321, *etc.*).

Similarly where components (e.g. modules and/or enzymatic domains) of the *blm* cluster are used to express various biomolecules (e.g. polyketides, sugars, polypeptides, *etc.*) the desired product and/or shunt metabolite(s) are isolated according to standard methods well known to those of skill in the art (see, e.g., Carreras and Khosla (1998) *supra*.) Purification and in vitro reconstitution of the essential protein components of an aromatic polyketide synthase. *Biochemistry* 37: 2084-2088, Deutscher (1990) *Methods in Enzymology Volume 182: Guide to Protein Purification*, M. Deutscher, ed. .

III. Synthesis of recombinant bleomycins.

In one embodiment this invention provides methods of synthesizing bleomycins and recombinantly synthesized bleomycins. As indicated above, this is generally accomplished by providing an organism (e.g. a bacterial cell) containing sufficient components of the *blm* gene cluster to direct synthesis of a complete bleomycin.

In one embodiment, the entire *blm* cluster is cloned into a *Streptomyces* strain (e.g., *S. lividans* or *S. coelicolor*). Kao *et al.* (1994) *Science*, 265: 509-512, have cloned the 30 kb DEBS genes from *Sacc. erythraea* into *S. coelicolor* and produced 6-deoxyerythronolide B in *S. coelicolor* and these methods can be used construct an expression plasmid for heterologous expression of the *blm* cluster. This method involves the transfer of DNA between a temperature-sensitive plasmid and a shuttle vector by means of a homologous double recombination event in *E. coli* (*Id.*). In a preferred embodiment, the two ends spanning the *blm* cluster are cloned into a temperature-sensitive plasmid that is chloramphenicol resistant (Cm^R) such as pCK6. *S. verticillus* DNA is then rescued from a donor into the temperature-sensitive recipient by co-transforming *E. coli* with the Cm^R recipient plasmid and the apramycin resistant (Ap^R) pKC505 donor cosmid that contains the *blm* gene cluster, followed by chloramphenicol and apramycin selection at 30°C. Colonies harboring both plasmids (Cm^R, Ap^R) will be shifted to 44°C on chloramphenicol and apramycin plates and only those cointegrates formed by a single recombination event

between the two plasmids are viable. Surviving colonies are then propagated at 30°C on Cm^R plates to select for recombinant plasmids formed by the resolution of cointegrates through a second recombinant event. The desired *blm* cluster is cloned into the Cm^R temperature-sensitive plasmid and is ready to be moved into any expression plasmid by a similar means of homologous recombinant event.

For example, if pWHM861 is the choice of shuttle plasmid for the expression of the *blm* cluster in *S. lividans* (Meurer and Hutchinson (1995) *J. Bacteriol.*, 177: 477-481), the two ends spanning the *blm* cluster downstream of the *ErmE** promoter in the ampicillin resistant (AM^R) plasmid pWHM861 are cloned. The resulting plasmid is co-transformed with the temperature-sensitive plasmid containing the *blm* cluster described above into *E. coli* under the selection of chloramphenicol and ampicillin at 30°C. These Cm^R and AM^R colonies are shifted to 44°C on chloramphenicol and ampicillin plates to undergo a single recombination event and the surviving colonies are resolved on ampicillin plates at 30°C by completing the double recombination process. The resulting plasmid is suitable for transformation into *S. lividans* by selection of thiostrepton, in which the expression of the desired *blm* cluster is under the control of the *ErmE** promoter. The *S. lividans* transformants are cultured and any metabolites produced are isolated and characterized.

Once production of BLM in *S. lividans* is established, mutated alleles of the *blm* synthetase can be introduced into the *blm* cluster for the production of BLM analogs.

IV. Altered endogenous expression of bleomycins.

Using the BLM gene cluster information provided herein, one of skill in the art may regulating the synthesis of endogenous bleomycin. The expression of various ORFs comprising the *blm* gene cluster may be increased or decreased to alter bleomycin synthesis levels.

Methods of altering the expression of endogenous genes are well known to those of skill in the art. Typically such methods involve altering or replacing all or a portion of the regulatory sequences controlling expression of the particular gene that is to be regulated. In a preferred embodiment, the regulatory sequences (*e.g.*, the native promoter) upstream of one or more of the *blm* ORFs are altered.

This is typically accomplished by the use of homologous recombination to introduce a heterologous nucleic acid into the native regulatory sequences. To downregulate expression of one or more *blm* ORFs, simple mutations that either alter the reading frame or disrupt the promoter are suitable. To upregulate expression of the *blm* ORF(s) the native

promoter(s) can be substituted with heterologous promoter(s) that induce higher than normal levels of transcription.

In a particularly preferred embodiment, nucleic acid sequences comprising the structural gene in question or upstream sequences are utilized for targeting heterologous recombination constructs.

The use of homologous recombination to alter expression of endogenous genes is described in detail in U.S. Patent 5,272,071, WO 91/09955, WO 93/09222, WO 96/29411, WO 95/31560, and WO 91/12650.

V. Synthesis of BLM analogs.

In one embodiment, this invention provides methods of synthesizing modified bleomycins or bleomycin analogs. In preferred embodiments, the BLM analogs are synthesized either by introducing specific perturbations into individual NRPS and/or PKS enzymatic domains or modules, or by reprogramming the linear order in which the NRPS or PKS enzymatic domains and/or modules appear in the *blm* synthetase genes. The former will lead to BLM analogs with targeted modifications at the BLM backbone and the latter will allow incorporation of other extension units in variable sequence into the biosynthesis of BLM. In particularly preferred embodiments, the genetically modified *blm* synthetases are produced in *S. verticillus*, however, it will be recognized that the entire *blm* gene cluster can be cloned into other hosts, e.g. into *S. lividans* or *S. coelicolor*.

In preferred embodiments modification of the *blm* gene cluster to yield BLM analogues is accomplished by one of two different approaches. In one approach, the BLM enzymatic domains and/or modules are altered in a directed manner (i.e. they are changed in a preselected way), while in another approach, random/haphazard alterations are introduced into the *blm* cluster and the resulting products are screened to identify those with desired properties.

A) Synthesis of BLM analogs by specific engineering of the *blm* synthetase genes.

The *blm* synthetase genes can be re-engineered by means of specific mutations or by reprogramming the linear order of the NRPS or PKS enzymatic domains or modules. In this approach, a wild-type *blm* synthetase allele is replaced with these mutants in and expressed in an appropriate host (e.g., *S. verticillus* or in a heterologous host). Since both NRPSs (Stachelhaus *et al.* (1995) *Science*, 269: 69-72) and PKSs (Donadio *et al.* (1993)

Proc. Natl. Acad. Sci. USA, 90: 7119-7123, Donadio *et al.* (1995) *J. Am., Chem. Soc.*, 117: 9105-9106, Cortes *et al.* (1995) *Science*, 268: 1487-1489) have shown considerable tolerance to reprogramming, it is expected that these modifications of the BLM synthetase will result in the production of BLM analogs with predicted structural alterations. For example,

5 targeted modification at the (2S,3S,4R)-4-amino-3-hydroxy-2-methylpentanoic acid AHM moiety of BLM can be accomplished by introduction of mutations into the *BLMVIII* PKS module of the BLM synthetase locus. Inactivation of the MT or KR motif by in-frame deletion or site-directed mutagenesis will result in the production of BLM analogs containing a demethyl-AHM, oxo-AHM, or oxo-demethyl-AHM moiety, *etc.*

10 Alternatively, individual functional NRPS domains and/or the PKS module can be deleted or the PKS module can be duplicated in-frame to produce BLM analogs with shorter or longer backbone, respectively. Alternatively, or in addition, the NRPS domains or the PKS module can be rearranged for the production of BLM analogs with a completely different backbone. The NRPS and PKS features can be combined into one integrated

15 system, providing access to a structural variation not available by either the NRPS or PKS system alone.

To create such mutations, plasmids are constructed carrying in-frame deletions of DNA segments encompassing a portion of the *blm* synthetase activities. Construction of specific deletions is preferably accomplished by one of the following two

20 strategies. The first involves subcloning of a DNA fragment in a gene replacement vector, selection of two restriction sites suitably located at the two ends of the DNA segments, and deletion of this segment from within the plasmid by rejoining the two resulting ends. An in-frame deletion can be obtained by a suitable combination of Klenow filling and S1 treatment of both ends prior to ligation.

25 The second approach involves polymerase chain reaction (PCR) amplification of two DNA segments that separate the region to be deleted followed by joining of the two fragments in the correct orientation in a gene replacement vector. This can be accomplished by designing PCR primers with suitable restriction sites. The restriction site used to generate the deletion and the sequences to serve as templates for the PCR amplification are chosen so

30 as to generate two segments of *blm* synthetase DNA of approximately equal length in the construction in order to maximize the chance of gene replacement. The gene replacement vector containing the allelic or deletion mutation is introduced into a *Streptomyces* strain (*e.g.*, *S. verticillus*). Integration of the plasmid into the *S. verticillus* chromosome via a single reciprocal homologous recombination will yield a recombinant that will be isolated by

selection for the vector marker. The resulting integrants are then grown under non-selective conditions and further resolution by selection for the loss of the vector marker via the second homologous recombination event will produce the desired deletion mutants.

- 5 Southern analysis of the isolated deletion mutants with the target DNA is performed to ensure that the expected double crossover recombination event has taken place. The first approach is convenient if there are suitably spaced restriction sites in the DNA sequence. The second approach enables the deletion of any DNA segment but may be limited by the size of the DNA segments that can be amplified by PCR. These *S. verticillius* recombinants are cultured under typical conditions for BLM production and the fermentation
10 broth is screened for the production of any novel BLM analogs resulted from the specific mutations in the *blm* synthetase locus.

B) Synthesis of BLM analogs by "random" modification of *blm* synthetase genes.

- 15 Bleomycin analogs can also be synthesized by randomly/haphazardly altering genes in the BLM cluster expressing the products of the randomly modified megasynthetase and then screening the products for the desired activity. Methods of "randomly" altering *blm* cluster genes are described below.

VI. Generation of other synthetic systems.

- 20 In addition to the production of bleomycin or modified bleomycins, the *blm* gene cluster or elements thereof can be used by themselves or in combination with NRPS and/or PKS modules and/or enzymatic domains of other PKS and/or NRPS systems to produce a wide variety of compounds including, but not limited to various polyketides, polypeptides, polyketide/polypeptide hybrids, various oxazoles and thiazoles, various sugars, various methylated polypeptides/polyketides, and the like. As with the production of
25 modified bleomycins described above, such compounds can be produced, *in vivo* or *in vitro*, by catalytic biosynthesis using large, modular PKSs, NRPSs, and hybrid PKS/NRPS systems. The megasynthetases directing such syntheses can be rationally designed *e.g.* by predetermined alteration/modification of polyketide and/or polypeptide and/or hybrid PKS/NRPS pathways. Alternatively, large combinatorial libraries of cells harboring various
30 megasynthetases can be produced by the random modification of particular pathways and then selected for the production of a molecule or molecules of interest. It will be appreciated that, in certain embodiments, such libraries of megasynthetases/modified pathways, can be

used to generate large, complex combinatorial libraries of compounds which themselves can be screened for a desired activity.

A) Directed modification of biomolecules.

Elements (*e.g.* open reading frames) of the *blm* biosynthetic gene cluster and/or variants thereof can be used in a wide variety of "directed" biosynthetic processes (*i.e.* where the process is designed to modify and/or synthesize one or more particular preselected metabolite(s)). Polypeptides encoded by particular open reading frames or combinations of open reading frames can be utilized to perform particular chemical modifications of biological molecules.

Thus, for example, open reading frames encoding a polypeptide synthetase can be used to chemically modify an amino acid by coupling it to another amino acid. In another example, the methyl transferase in *BlmVIII* can be utilized to introduce methyl groups into polyketides, and other, substrates. The glycosyl transferases can be used to glycosylate appropriate substrates, and so forth. These examples, are merely illustrative. One of skill in the art, utilizing the information provided here, can perform literally countless chemical modifications and/or syntheses using either "native" bleomycin biosynthesis metabolites as the substrate molecule, or other molecules capable of acting as substrates for the particular enzymes in question. Other substrates can be identified by routine screening. Methods of screening enzymes for specific activity against particular substrates are well known to those of skill in the art.

The biosyntheses can be performed *in vivo*, *e.g.* by providing a host cell comprising the desired *blm* gene cluster open reading frame(s) and/or *in vivo*, *e.g.*, by providing the polypeptides encoded by the *blm* gene cluster ORFs and the appropriate substrates and/or cofactors.

B) Directed engineering of novel synthetic pathways.

In numerous embodiments of this invention, novel polyketides, polypeptides, and combinations thereof are created by modifying known PKSs or NRPSs so as to introduce variations into known polymers synthesized by the enzymes. Such variations may be introduced by design, for example to modify a known molecule in a specific way, *e.g.* by replacing a single monomeric unit within a polymer with another, thereby creating a derivative molecule of predicted structure. Such variations can also be made by adding one or more modules to a known PKS or NRPS, or by removing one or more module from a

known PKS or NRPS. Such novel PKSs or NRPSs can readily be made using a variety of techniques, including recombinant methods and *in vitro* synthetic methods.

Using any of these methods, it is possible to introduce PKS domains into a NRPS, or vice versa, thereby creating novel molecules including both peptide and polyketide structural domains. For example, a PKS enzyme producing a known polyketide can be modified so as to include an additional module that adds a peptide moiety into the polyketide. Novel molecules synthesized using these methods can be screened, using standard methods, for any activity of interest, such as antibiotic activity, effects on the cell cycle, effects on the cytoskeleton, etc.

Novel polyketides, polypeptides, or combinations thereof can also be made by creating novel PKSs or NRPSs *de novo*, using recombinant or *in vitro* synthetic methods. Such novel arrangements of domains can be designed, *i.e.* to create a specific polymer. In addition to creating novel PKSs or NRPSs by combining modules, the methods of this invention can also be used to make novel modules that can add new monomeric units to a growing polypeptide or polyketide chain. Because the identity of each module, and, consequently, the identity of the monomer added by the module, is determined by the identity and number of the functional domains comprising the module, it is possible to produce novel monomeric units by creating novel combinations of functional domains within a module. Such novel modules can be created by design, for example to make a specific module that will add a specific monomer to a polyketide or polypeptide, or can be created by the random association of domains so as to produce libraries of novel modules. Such novel modules can be made using recombinant or *in vitro* synthetic means.

Mutations can be made to the native NRPS and/or PKS subunit sequences and such mutants used in place of the native sequence, so long as the mutants are able to function with other PKS and/or PKS subunits to collectively catalyze the synthesis of an identifiable polyketide and/or polypeptide. Such mutations can be made to the native sequences using conventional techniques such as by preparing synthetic oligonucleotides including the mutations and inserting the mutated sequence into the gene encoding a NRPS and/or PKS subunit using restriction endonuclease digestion. (*see, e.g., Kunkel, (1985) Proc. Natl. Acad. Sci. USA 82: 448; Geisselsoder et al. (1987) BioTechniques 5: 786*). Alternatively, the mutations can be effected using a mismatched primer (generally 10-20 nucleotides in length) which hybridizes to the native nucleotide sequence, at a temperature below the melting temperature of the mismatched duplex. The primer can be made specific by keeping primer length and base composition within relatively narrow limits and by keeping the mutant base

centrally located (Zoller and Smith (1983) *Meth. Enzymol.* 100: 468). Primer extension is effected using DNA polymerase, the product cloned and clones containing the mutated DNA, derived by segregation of the primer extended strand, selected. Selection can be accomplished using the mutant primer as a hybridization probe. The technique is also applicable for generating multiple point mutations (see, e.g., Dalbie-McFarland *et al.* (1982) *Proc. Natl. Acad. Sci USA* 79:6409). PCR mutagenesis will also find use for effecting the desired mutations.

C) Random modification of PKS/NRPS pathways.

In another embodiment, variations can be made randomly, for example by making a library of molecular variants of a known polymer by randomly mutating one or more PKS or NRPS modules and/or enzymatic domains or by randomly replacing one or more modules or enzymatic domains in a known PKS or NRPS with a collection of alternative modules and/or enzymatic domains..

The PKS and/or NRPS modules can be combined into a single multi-modular enzyme, thereby dramatically increasing the number of possible combinations obtained using these methods. These combinations can be made using standard recombinant or nucleic acid amplification methods, for example by shuffling nucleic acid sequences encoding various modules or enzymatic domains to create novel arrangements of the sequences, analogous to DNA shuffling methods described in Crameri *et al.*, (1998) *Nature* 391: 288-291, and in U.S. Patents 5,605,793 and in 5,837,458. In addition, novel combinations can be made *in vitro*, for example by combinatorial synthetic methods. Novel polymers, or polymer libraries, can be screened for any specific activity using standard methods.

Random mutagenesis of the nucleotide sequences obtained as described above can be accomplished by several different techniques known in the art, such as by altering sequences within restriction endonuclease sites, inserting an oligonucleotide linker randomly into a plasmid, by irradiation with X-rays or ultraviolet light, by incorporating incorrect nucleotides during *in vitro* DNA synthesis, by error-prone PCR mutagenesis, by preparing synthetic mutants or by damaging plasmid DNA *in vitro* with chemicals. Chemical mutagens include, for example, sodium bisulfite, nitrous acid, hydroxylamine, agents which damage or remove bases thereby preventing normal base-pairing such as hydrazine or formic acid, analogues of nucleotide precursors such as nitrosoguanidine, 5-bromouracil, 2-aminopurine, or acridine intercalating agents such as proflavine, acriflavine, quinacrine, and the like.

Generally, plasmid DNA or DNA fragments are treated with chemicals, transformed into *E. coli* and propagated as a pool or library of mutant plasmids.

Large populations of random enzyme variants can be constructed *in vivo* using "recombination-enhanced mutagenesis." This method employs two or more pools of, for example, 10^6 mutants each of the wild-type encoding nucleotide sequence that are generated using any convenient mutagenesis technique, described more fully above, and then inserted into cloning vectors.

D) Incorporation and/or modification of non-blm cluster elements.

In either the directed or random approaches, nucleic acids encoding novel combinations of modules and/or enzymatic are introduced into a cell. In one embodiment, nucleic acids encoding one or more PKS or NRPS domains are introduced into a cell so as to replace one or more domains of an endogenous PKS or NRPS within a chromosome of the cell. Endogenous gene replacement can be accomplished using standard methods, such as homologous recombination. Nucleic acids encoding an entire PKS, NRPS, or combination thereof can also be introduced into a cell so as to enable the cell to produce the novel enzyme, and, consequently, synthesize the novel polymer. In a preferred embodiment, such nucleic acids are introduced into the cell optionally along with a number of additional genes, together called a 'gene cluster,' that influence the expression of the genes, survival of the expressing cells, etc. In a particularly preferred embodiment, such cells do not have any other PKS- or NRPS- encoding genes or gene clusters, thereby allowing the straightforward isolation of the polymer synthesized by the genes introduced into the cell.

Furthermore, the recombinant vector(s) can include genes from a single PKS and/or NRPS gene cluster, or may comprise hybrid replacement PKS gene clusters with, e.g., a gene for one cluster replaced by the corresponding gene from another gene cluster. For example, it has been found that ACPs are readily interchangeable among different synthases without an effect on product structure. Furthermore, a given KR can recognize and reduce polyketide chains of different chain lengths. Accordingly, these genes are freely interchangeable in the constructs described herein. Thus, the replacement clusters of the present invention can be derived from any combination of PKS and/or NRPS gene sets that ultimately function to produce an identifiable polyketide and/or peptide.

Examples of hybrid replacement clusters include, but are not limited to, clusters with genes derived from two or more of the *act* gene cluster, the *whiE* gene cluster, frenolicin (*fren*), granaticin (*gra*), tetracenomyacin (*tcm*), 6-methylsalicylic acid (6-msas),

oxytetracycline (*otc*), tetracycline (*tet*), erythromycin (*ery*), griseusin (*gris*), nanaomycin, medermycin, daunorubicin, tylosin, carbomycin, spiramycin, avermectin, monensin, nonactin, curamycin, rifamycin and candididin synthase gene clusters, among others. (For a discussion of various PKSs, see, e.g., Hopwood and Sherman (1990) *Ann. Rev. Genet.* 24:

5 37-66; O'Hagan (1991) *The Polyketide Metabolites*, Ellis Horwood Limited.

A number of hybrid gene clusters have been constructed, having components derived from the *act*, *fren*, *tcm*, *gris* and *gra* gene clusters (see, e.g., U.S. Patent 5,712,146). Other hybrid gene clusters, as described above, can easily be produced and screened using the disclosure herein, for the production of identifiable polyketides, polypeptides or
10 polyketide/polypeptide hybrids.

Host cells (e.g. *Streptomyces*) can be transformed with one or more vectors, collectively encoding a functional PKS/NRPS set (e.g. a bleomycin or bleomycin analog), or a cocktail comprising a random assortment of PKS and/or NRPS genes, modules, active sites, or portions thereof. The vector(s) can include native or hybrid combinations of PKS
15 and/or NRPS subunits or cocktail components, or mutants thereof. As explained above, the gene cluster need not correspond to the complete native gene cluster but need only encode the necessary PKS and/or NRPS components to catalyze the production of the desired product. For example, in *Streptomyces* aromatic PKSs, carbon chain assembly requires the products of three open reading frames (ORFs). ORF1 encodes a ketosynthase (KS) and an
20 acyltransferase (AT) active site (KS/AT); ORF2 encodes a chain length determining factor (CLF), a protein similar to the ORF1 product but lacking the KS and AT motifs; and ORF3 encodes a discrete acyl carrier protein (ACP). Some gene clusters also code for a ketoreductase (KR) and a cyclase, involved in cyclization of the nascent polyketide backbone. However, it has been found that only the KS/AT, CLF, and ACP, need be present
25 in order to produce an identifiable polyketide. Thus, in the case of aromatic PKSs derived from *Streptomyces*, these three genes, without the other components of the native clusters, can be included in one or more recombinant vectors, to constitute a "minimal" replacement PKS gene cluster.

E) Variation of starter and extender units.

30 In addition to varying the PKS and/or NRPS modules and/or domains, variations in the products produced by various PKS/NRPS systems can be obtained by varying the starter units and/or the extender units. Thus, for example, a considerable degree of variability exists for starter units, e.g., acetyl CoA, malonyl CoA, propionyl CoA,

acetate, butyrate, isobutyrate and the like. In addition, naturally occurring PKSs and/or NRPSs have shown some tolerance for varying extender units.

F) Examples of preferred modifications.

As indicated above, the novel PKS and NRPS modules and enzymatic domains identified herein can be used to perform specific single modifications of particular substrates, or as components of complex synthetic pathways to generate particular products or large combinatorial libraries. As described in the Examples, a number of modules of the *blm* gene cluster provide novel functionality. By way of example, a few preferred reactions are listed below. These examples are intended to be illustrative and are not exhaustive nor limiting.

1. Use of *BlmVIII* PKS to introduce branched methyl group.

The *blmVIII* gene identified herein encodes a PKS module consisting of domains characteristic for known PKSs, such as ketoacyl synthase (KS), acyltransferase (AT), ketoreductase (KR), and ACP, with malonyl CoA acting as an extending unit. However, the identification of an integrated methyltransferase (MT) domain in the middle of *BlmVIII* is unique, representing the first PKS from actinomycetes that contains an internal MT domain. The use of this methyltransferase domain allows the introduction of a branched methyl group during a polyketide and/or polypeptide and/or hybridizing polyketide/polypeptide synthesis. Figure 5 illustrates the use of *BlmVIII* PKS in engineering a polyketide biosynthesis that introduces a branched methyl group.

The first formula in Figure 5 illustrates a polyketide synthesis mediated by 6-deoxyerythronolide B synthase (DEBS) which normally catalyzes the biosynthesis of the erythromycin aglycone, 6-deoxyerythronolide B. The remaining formulas show how the use of the *blmVIII* methyltransferase (MT) group at different points in the synthesis results in the introduction of a methyl group at different locations in the resulting product.

In view of this illustration, one of skill in the art would appreciate that the *blmVIII* MT domain can be used in a wide variety of biosyntheses to introduce methyl branches.

2. Use of the *blm* gene cluster to make thiazolidine, thiazoline, thiazole, bi-thiazolidine, bithiazoline, and bithiazole-containing compounds.

The *BlmIV* and *BlmIII* NRPSs are characterized by unusual Cy domains as well as an unprecedented Ox domain, providing an efficient biosynthesis for a bithiazole structure. While thiazoline is the direct product of the Cy domain, the thiazoline-to-thiazole conversion generally is performed with an additional oxidation step. We identified at the C-terminus of NRPS-0 an additional domain that shows low, but significant, sequence homology to a family of putative oxidases/dehydrogenases, including the McbC protein of the microcin B17 synthase (Table 1). Microcin B17 synthase catalyzes the synthesis of the oxazole and thiazole-containing peptide antibiotic microcin B17, and McbC has been proposed to play a role in catalyzing the oxazoline/thiazoline-to-oxazole/thiazole conversion. Consequently, we propose that this extra domain at the C-terminus of NRPS-0 provides the oxidase/dehydrogenase activity for the biosynthesis of the bithiazole moiety of BLM, defining a novel Ox domain for NRPSs.

It is noteworthy that a cell-free preparation from *Sv* ATCC15003 has been reported to catalyze the conversion of phleomycins to BLMs in the presence of NAD^+ , supporting the hypothesis that the bithiazole moiety of BLM results from stepwise oxidations of a bithiazoline precursor (Fig. 1A). (The phleomycin producer could be imagined to result from the loss of its Ox activity for the first thiazoline ring.) Given the wide distribution of thiazole or oxazole rings in natural products exhibiting an impressive array of biological activities, the cloning of the *blmIV*, *III* genes and the identification of the Ox domain open many opportunities thiazole biosynthesis and to synthesize novel thiazole containing molecules by engineering peptide biosynthesis.

Representative thiazole syntheses using variants of the *blm* NRPS are illustrated in Figure 6. Note that in Figure 6, A^{M} and A^{N} refer to an A domain that activates and amino acid with R^{M} and R^{N} groups, respectively. A^{C} refers to an A domain that activates Cys ($\text{x} = \text{SH}$) or Ser ($\text{X} = \text{OH}$) that can be cyclized to form the oxiazoline/thiazoline or oxazole/thiazole structures. DH is a dehydratase. In view of these representative examples, one of skill in the art would appreciate that the *blm* NRPS domain and its variants can be used in a wide variety of chemical syntheses make thiazolidine, thiazoline, thiazole, bi-thiazolidine, bithiazoline, or bithiazole-containing compounds.

3. Use of the *blm* gene cluster to make heterocyclic ring-containing compounds.

Various *blm* modules can be used to produce heterocyclic ring-containing compounds. Such heterocycles include, but are not limited to five member S- and N-containing compounds of the thiazolidine, thiazoline and thiazole family or the O- and N-containing compounds of the oxazolidine, oxazoline, and oxazole family. Again, the preparation of such compounds is illustrated in Figure 6.

4. Use of the *blm* gene cluster to make sugars.

In still another embodiment, the *blm* gene cluster or elements thereof can be used to make sugars. Such sugars include, but are not limited to L-sugars (with the *BlmG* epimerase), sugars modified by a carbamoyl group (*e.g.*, using *BlmD*), and various disaccharides. Representative examples of such syntheses are illustrated in Figure 7. Such sugar biosynthesis genes can also be used to attach sugars onto other polyketide and/or peptide aglycones.

F) Screening of products.

Particularly where large combinatorial libraries are synthesized, *e.g.* using one or more modules and/or enzymatic domains of the *blm* gene cluster it will often be desired to screen the resulting compound(s) for the desired activity. Methods of screening compounds (*e.g.* polypeptides, polyketides, sugars, thiazoles, *etc.*) for various activities of interest (*e.g.* cytotoxicity, antimicrobial activity, particular chemical activities, *etc.*) are well known to those of skill in the art.

Where large numbers of compounds are produced, it is often desired to rapidly screen such compounds using "high throughput systems" (HTS). High throughput assays systems are well known to those of skill in the art and many such systems are commercially available. (*see, e.g.*, Zymark Corp., Hopkinton, MA; Air Technical Industries, Mentor, OH; Beckman Instruments, Inc. Fullerton, CA; Precision Systems, Inc., Natick, MA, *etc.*). These systems typically automate entire procedures including all sample and reagent pipetting, liquid dispensing, timed incubations, and final readings of the microplate in detector(s) appropriate for the assay. These configurable systems provide high throughput and rapid start up as well as a high degree of flexibility and customization. The manufacturers of such systems typically provide detailed protocols for the various high throughput screens.

VII. In Vitro syntheses.

In additional embodiments of this invention, bleomycins and other polyketides and/or polypeptides are synthesized and/or modified *in vitro*. Individual enzymatic domains or modules can be used *in vitro* to modify a unit and/or to add a single monomeric unit to a growing polyketide or polypeptide chain. In one approach a metatransferase providing all the desired synthetic activities recombinantly expressed and then provided, the appropriate substrates and buffer system *e.g.* in a bioreactor, to direct the synthesis of the desired product. In another approach, various PKSs and/or NRPSs are provided in different solutions and the growing polymer chains can be sequentially introduced into the plurality of solutions, each containing a single (or several) PKS or NRPS modules. In still another embodiment, the PKS and/or NRPS modules or enzymatic domains are provided attached to a solid support and a fluid containing the growing macromolecule is passed over the surface whereby the PKSs or NRPSs are able to react with the target substrate.

In one preferred embodiment, a combinatorial library of polyketides or polypeptides, or combinations thereof, is created by using automated means to facilitate the sequential introduction of a multitude of polymeric chains, each attached to a solid support, to a collection of solutions, each containing a single PKS or NRPS module. These automated means can be used to systematically vary the sequence by which each polymeric chain is introduced into the various solutions, thereby creating a combinatorial library. Numerous methods are well known in the art to create combinatorial libraries of molecules by the sequential addition of monomeric units, for example as described in WO 97/02358.

VIII. Kits.

In still another embodiment, this invention provides kits for practice of the methods described herein. In one preferred embodiment, the kits comprise one or more containers containing nucleic acids encoding one or more of the *blm* gene cluster ORFs and/or one or more of the BLM PKS or NRPS modules or enzymatic domains. Certain kits may comprise vectors encoding the *blm* orfs and/or cells containing such vectors. The kits may optionally include any reagents and/or apparatus to facilitate practice of the assays described herein. Such reagents include, but are not limited to buffers, labels, labeled antibodies, bioreactors, cells, *etc.*

In addition, the kits may include instructional materials containing directions (*i.e.*, protocols) for the practice of the methods of this invention. Preferred instructional

materials provide protocols utilizing the kit contents for creating or modifying *blm* module or ORF and/or for synthesizing or modifying a molecule using one or more *blm* modules and/or enzymatic domains. While the instructional materials typically comprise written or printed materials they are not limited to such. Any medium capable of storing such instructions and communicating them to an end user is contemplated by this invention. Such media include, but are not limited to electronic storage media (e.g., magnetic discs, tapes, cartridges, chips), optical media (e.g., CD ROM), and the like. Such media may include addresses to internet sites that provide such instructional materials.

EXAMPLES

The following examples are offered to illustrate, but not to limit the claimed invention.

Example 1

Bleomycin biosynthesis in *Streptomyces verticillus* ATCC15003, A model for hybrid peptide and polyketide biosynthesis.

Here we report the cloning and characterization of the *blm* biosynthesis gene cluster from *Sv* ATCC15003 (Fig. 2). Sequence analysis and biochemical characterization of individual modules enabled us to align the nine NRPS and one PKS modules in a linear order to constitute the Blm megasynthetase complex (Fig. 1B). These studies revealed several unprecedented features for peptide and polyketide biosynthesis, setting the stage to investigate the molecular basis for intermodular communication between NRPS and PKS, and supported the wisdom of combining individual NRPS and PKS modules for combinatorial biosynthesis to make novel “unnatural” natural products from amino acids and short carboxylic acids.

Materials and Methods.

General procedures.

Escherichia coli DH5 α (Sambrook *et al.* (1989) *Molecular Cloning: A Laboratory Manual*, 2nd ed, Cold Spring Harbor Laboratory Press, Cold Spring Harbor, USA), *E. coli* XL 1-Blue MR (Stratagene, La Jolla, CA), *E. coli* BL21(DE-3) (Novagen, Madison, WI), and *Sv* ATCC15003 (American Type Culture Collection, Rockville, MD) were used in this work. pOJ446 (Agricultural Research Service Culture Collection, Peoria, IL), pQE60 (Qiagen, Santa Clarita, CA), pET28a and pET29a (Novagen), and other plasmids

were from commercial sources. *E. coli* (Sambrook, *supra.*) and *Sv* ATCC15003 strains (Hopwood *et al.* (1985) *Genetic Manipulation of Streptomyces: A Laboratory Manual*, The John Innes Foundation, Norwich, UK) were cultured under standard conditions.

- Plasmid preparation was carried out by using commercial kits (Qiagen). Total
- 5 *Sv* ATCC15003 DNA was isolated according to literature protocols (Hopwood *et al.* (1985) *Genetic Manipulation of Streptomyces: A Laboratory Manual*, The John Innes Foundation, Norwich, UK; Nagaraja *et al.* (1987) *Methods Enzymol.* 153: 166-198). Restriction enzymes and other molecular biology reagents were from commercial sources, and digestions and ligation followed standard methods (Sambrook, *supra.*). For Southern analysis, digoxigenin
- 10 labelling of DNA probes, hybridization, and detection were performed according to the protocols provided by the manufacturer (Boehringer Mannheim Biochemicals, Indianapolis, IN).

- Automated DNA sequencing was carried out on an ABI Prism 377 DNA Sequencer (Perkin-Elmer/ABI, Foster City, CA), and this service was provided by either the
- 15 DBS Automated DNA Sequencing Facility, UC Davis, or Davis Sequencing (Davis, CA). Data were analyzed by the ABI Prism Sequencing 2.1.1 software and the Genetics Computer Group (GCG) program (Madison, WI).

Cloning and sequencing of the *blm* gene cluster.

- A genomic library of *Sv* ATCC15003 was constructed in pOJ446 according to
- 20 literature procedures (Nagaraja *et al.* (1987) *Methods Enzymol.* 153: 166-198) and screened with probes made from both ends of the *blmAB* locus (Sugiyama *et al.* (1994) *Gene* 151: 11-16; Calcutt and Schmidt (1994) *Gene* 151: 17-21), leading to the localization of 140-kb contiguous DNA, of which 100-kb is upstream (Fig. 2) and 40-kb is downstream (data not shown) of the *blmAB* genes. Heterologous NRPS probes were amplified from *Sv*
- 25 ATCC15003 by polymerase chain reaction (PCR) according to literature procedures (Turgay and Marahiel (1994) *Peptide Res.* 7: 238-241) and used to screen the entire 140-kb DNA by Southern analysis under various hybridization conditions (Shen *et al.* (1999) *Bioorg. Chem.* 27: 155-171).

Prediction of substrate specificity of NRPSs.

- 30 The nine Blm NRPS modules were compared with eighty four modules from various bacterial and fungal NRPSs available at the GenBank, including those with known or putative specificity for amino acids present in BLM. A table of overall similarities/identities

was generated by PILEUP analysis of the A3 to A6 regions, and the residues lining the substrate binding pocket by comparison with PheA (Conti *et al.* (1997) *EMBO J.* 16, 4174-4183) were determined by PILEUP/PRETTY analysis. The percentage similarities for each Blm NRPS module were plotted against the rest of the NRPS modules to display the overall sequence homology between the A3 to A6 region. Those modules that showed significantly higher homology were selected to compare the amino acid residues that line the substrate binding pocket.

Overproduction and biochemical characterization of the NRPS-1A and NRPS-6A proteins.

Heterologous expression of the A domain in *E. coli* were performed according to literature procedures (Mootz and Marahiel (1997) *J. Bacteriol.* 179: 6843-6850). NRPS-1A (forward primer 5'-AAC CCA TGG CTG CTT CCC TGA CCC GCC TGG CC-3', SEQ ID NO:76, and reverse primer 5'-CCT AGA TCT ACG GGC AGG TGG GGC GGT-3', SEQ ID NO:77) and NRPS-6A (forward primer 5'-GGG AAT TCC ATA TGA TCC TCA CGT CCT TCC AC-3', SEQ ID NO:78, and reverse primer 5'-GGC AAG CTT GGG TGA GGG TCC GTT CGG T-3', SEQ ID NO:79) were amplified by PCR from *Sv* ATCC15003 cosmid clones. The resulting 1.6-kb fragment of NRPS-1A was first cloned into the *NcoI/BglIII* sites of pQE60 and then moved as an *NcoI/HindIII* fragment into the similar sites of pET29a to yield pBS10, and the resulting 1.6-kb fragment of NRPS-6A was directly cloned into the *NdeI/HindIII* sites of pET28a to yield pBS11. Introduction of pBS10 and pBS11 into *E. coli* BL21(DE-3) under standard expression conditions resulted in production of NRPS-1A (with an N-terminal S-tag and a C-terminal His₆-tag) and NRPS-6A (with an N-terminal His₆-tag), respectively. The soluble fractions of fusion proteins were subjected sequentially to an affinity chromatography on Ni-NTA resin and an anion exchange chromatography on a Hyper-D column (PerSeptive Biosystem, Framingham, MA), resulting in NRPS-1A and NRPS-6A with near homogeneity.

Results and Discussion.

Cloning of the *blm* gene cluster from *Sv* ATCC15003.

Davies and co-workers previously cloned two BLM resistance genes (*blmA* and *blmB*) from *Sv* ATCC15003 (Sugiyama *et al.* (1994) *Gene* 151: 11-16), and Calcutt and Schmidt (1994) *Gene*, 151: 17-21, sequenced a 7.2-kb DNA fragment flanking the *blmAB*

genes, revealing seven open reading frames (orfs), none of which were found to encode Blm NRPS or PKS enzymes. Given the precedent that antibiotic production genes commonly occur as a cluster in actinomycetes, we adopted an approach combining chromosomal walking from the *blmAB* resistance locus and DNA hybridization with heterologous NRPS probes to clone and identify the *blm* cluster, leading to the localization of 140-kb contiguous *Sv* ATCC15003 DNA. DNA sequencing of approximately 90-kb of the *blm* gene cluster, including the 7.2-kb *blmAB* locus, revealed 40 ORFs (Fig. 2). Preliminary functional assignments were made by comparison of the deduced gene products with proteins of known functions in the database. Among the ORFs identified from the *blm* cluster, we indeed found a PKS module, flanked by several NRPS modules—a fact that supports the hybrid NRPS/PKS/NRPS hypothesis for BLM biosynthesis—along with several sugar biosynthesis genes and genes encoding other biosynthesis enzymes as well as several resistance and regulatory genes (Table 1).

Noteworthy are the genes encoding the putative NRPS and PKS enzymes.

The *blmI*, *blmII*, and *blmXI* genes encode NRPSs with an unusual architecture. In contrast to all known NRPSs, which are of modular organization with each module consisting minimally of a condensation (C), an adenylation (A), and a peptidyl carrier protein (PCP) domain (1), BlmI, BlmII, and BlmXI are discrete proteins homologous to individual domains of type I NRPSs. We have characterized BlmI as a type II PCP (18). The BlmII and BlmXI proteins could serve as candidates for type II condensation enzymes. It is unclear yet what role if any these discrete NRPS enzymes could play in BLM biosynthesis.

The *blmIII*, *blmIV*, *blmV*, *blmVI*, *blmVII*, *blmIX*, and *blmX* genes encode modular NRPSs consisting of domains characteristic for known type I NRPSs (A special thematic issue on polyketide and nonribosomal polypeptide biosynthesis, (1997) *Chem. Rev.* 97: 2463-2706), such as the A, PCP, C, and condensation/cyclization (Cy) domains (Konz *et al.* (1997) *Chem. Biol.* 4: 927-937), as well as an unprecedented oxidation (Ox) domain (see discussion below). However, BlmVI is unique among all the Blm NRPSs identified. Its N-terminal module (NRPS-5) consists of an atypical A domain, which bears a close resemblance to a family of acyl CoA synthases (Fitzmaurice and Kolattukudy (1997) *J. Bacteriol.* 179: 2608-2615; Fitzmaurice and Kolattukudy (1998) *J. Biol. Chem.* 273: 8033-8039), and an acyl carrier protein (ACP)-like domain (A special thematic issue on polyketide and nonribosomal polypeptide biosynthesis, (1997) *Chem. Rev.* 97: 2463-2706). Its C-terminal module is truncated and presumably interacts with BlmV to constitute the complete NRPS-3 module (Fig. 1B). Also noteworthy are the C domain of NRPS-3 that lacks both

His residues of the conserved HHxxxDG (SEQ ID NO:4) active site for transpeptidation (Stachelhaus *et al.* (1998) *J. Biol. Chem.*, 273: 22773-22781) and the extra C domain at the C-terminus of BlmV. These unusual features associated with *BlmVI* and *BlmV* may play roles in the formation of the β -aminoalaninamide and the pyrimidine moieties of BLM, which are unprecedented in peptide biosynthesis. For example, we propose that the NRPS-4-activated Ser is first dehydrated into dehydroalanine before condensation—an analogous Thr-to-2,3-dehydroaminobutyric acid dehydration has been observed in syringomycin biosynthesis (Guenzi *et al.* (1998) *J. Biol. Chem.* 273: 32857-32863). Conjugate addition to dehydroalanine by Asn on the NRPS-3 module downstream followed by an aminolysis to cleave the Ser-Asn adduct off the Blm megasynthetase furnishes the β -aminoalaninamide moiety (Fig. 1B). The former reaction could be catalyzed by the C domain of NRPS-3 that apparently is nonfunctional for normal transpeptidation due to the lack of the active sites, and the latter reaction could be catalyzed by the acyl CoA synthase-like domain of NRPS-5 in a process that resembles the acyl CoA synthase-catalyzed synthesis of acyl CoA from carboxylic acid (Stachelhaus *et al.* (1998) *J. Biol. Chem.* 273: 22773-22781; Guenzi *et al.* (1998) *J. Biol. Chem.* 273: 32857-32863) but in the reverse direction in the presence of an amino donor (Fig. 1B).

The *blmVIII* gene encodes a PKS module consisting of domains characteristic for known PKSs, such as ketoacyl synthase (KS), acyltransferase (AT), ketoreductase (KR), and ACP, with malonyl CoA acting as an extending unit according to sequence comparison of the AT domain (Haydock *et al.* (1995) *FEBS Lett.* 374: 246-248) (Fig. 1B). However, the identification of an integrated methyltransferase (MT) domain (Kagan and Clarke (1994) *Arch. Biochem. Biophys.* 310: 417-427) in the middle of *BlmVIII* is unique, representing the first PKS from actinomycetes that contains an internal MT domain. The only other example of PKS from bacteria that contains an internal MT domain is HMWP1 of the yersiniabactin gene cluster (Pelludat *et al.* (1998) *J. Bacteriol.* 180: 538-546). It has been assumed that fungal PKSs in general contain internal MTs for the introduction of methyl branch into the polyketide products, as it has been shown recently in lovastatin biosynthesis (Kennedy *et al.* (1999) *Science* 284: 1368-1372).

The Blm megasynthetase-templated assembly of BLM.

According to the hybrid NRPS/PKS/NRPS model for BLM biosynthesis (Fig. 1A), we predict a linear modular organization of individual NRPS and PKS modules to constitute the Blm megasynthetase. Thus, the first functional domain of the Blm

megasyntetase should be a NRPS module that initiates BLM biosynthesis by activating L-Ser as an amino acylthioester to set the stage for transpeptidation. Chain elongation proceeds by sequential incorporation of L-Asn, L-Asn, L-His, and L-Ala, requiring four additional NRPS modules. In the next step, a malonate reacts with the resulting pentapeptide intermediate to form a β -ketothioester intermediate that is subsequently methylated at the α -position and reduced at the β -keto group. A PKS module presumably dictates all these biosynthetic events and interacts with the aligned NRPS module upstream to channel the growing peptide intermediate from an NRPS module to a PKS module. After one cycle of polyketide elongation, peptide elongation is resumed by incorporation of an L-Thr residue. This step is presumably catalyzed by an NRPS module that interacts with the upstream PKS module to channel the growing polyketide intermediate (as far as the active site is concerned) from a PKS module to an NRPS module. At this stage, methylation occurs at the pyrimidine moiety of the growing intermediate, presumably catalyzed by a discrete methyltransferase; chain elongation is continued by three additional NRPS modules that incorporate a β -Ala and two L-Cys molecules sequentially. Finally, the fully assembled BLM peptide/polyketide/peptide backbone is hydroxylated at the β -position of the His residue, presumably by a discrete hydroxylase, and released from the Blm megasyntetase complex via nucleophilic substitution of the RCO-S-PCP species by a terminal amine to form the BLM aglycone. Intermediates after five of the nine proposed elongation steps were in fact isolated as P-3, P-3A, P-3K, P-4, P-5, P-5m, P-6m, and P-6mo (Takita and Muroka (1990) pages 289-309 in *Biochemistry of Peptide Antibiotics: Recent Advances in the Biotechnology of β -Lactams and Microbial Peptides*, Kleinkauf, H. & von Döhren, H. eds., W. de Gruyter, N.Y.), which presumably resulted from premature departure from the Blm megasyntetase complex before the chain reaches its full length (Fig. 1B).

Most of the bacterial NRPS gene clusters characterized to date are organized in operon-type structures, encoding multimodular NRPS proteins with individual modules organized along the chromosome in a linear order that parallels the order of the amino acids in the resultant peptides, i.e., following the “colinearity rule” for the NRPS-templated assembly of peptides from amino acids (A special thematic issue on polyketide and nonribosomal polypeptide biosynthesis, (1997) *Chem. Rev.* 97: 2463-2706; Cane *et al.* (1998) *Science* 282: 63-68). Inspection of the *blm* gene cluster (Fig. 2) showed that the Blm NRPS and PKS modules apparently are not organized according to the “colinearity rule” for BLM biosynthesis (Fig. 1). [Exception to the “colinearity rule” was also noted in the

syringomycin synthetase gene cluster (Guenzi *et al.* (1998) *J. Biol. Chem.* 273: 32857-32863), and in fact, Grandi and co-workers have demonstrated recently in *Bacillus subtilis* that neither the operon-type structure nor the physical linkage of individual modules is essential for proper assembly and activity of the surfactin NRPS megasynthetase (Guenzi *et al.* (1998) *J. Biol. Chem.* 273: 14403-14410).] Realizing that the BLM biosynthesis cannot be rationalized according to the “colinearity rule”, we determined the substrate specificity of individual NRPS and PKS modules in an attempt to shed light on the modular organization of the Blm megasynthetase complex. Brick and co-workers postulated, based on the X-ray structural analysis of the A domain of GrsA, PheA, that the region between core sequences A3 to A6 represent the amino acid specificity determinant of an NRPS module (Conti *et al.* (1997) *EMBO J.* 16: 4174-4183). Since the A domains in all known NRPSs share a significant sequence identity (ensuring that the main chain conformation of the enzymes is likely to be very similar), they further proposed that the differing substrate specificity of individual NRPS modules will be mainly determined by the nature of the amino acids lining the substrate binding pocket (Stachelhaus *et al.* (1999) *Chem. Biol.* 6: 493-505; Conti *et al.* (1997) *EMBO J.* 16: 4174-4183). Given this structural information and the vast amount of NRPS sequences available at the GenBank, we developed a novel approach for predicting substrate specificity for NRPS modules by comparing the overall sequence between the A3 to A6 region and the eight amino acid residues that line up the substrate binding pocket. While a constant level of similarities (30%-40%) was evident among all the NRPS modules analyzed, most of the Blm NRPS modules showed striking similarities (50%-60%) to a particular cluster of NRPS modules as exemplified in Fig. 3A for NRPS-1 and NRPS-6. Close examination of these modules clustered with higher similarities revealed that they activate the same or very similar amino acid, based on which the putative substrate for the NRPS in query could be predicted, i.e., NRPS-1 and NRPS-6A activate L-Cys and L-Thr, respectively. These predictions were further supported by comparing the residues lining the substrate binding pocket. For example, the amino acid residues lining the substrate binding pocket for NRPS-1 and NRPS-6 are almost identical to those NRPS modules that are known to activate L-Cys and L-Thr, respectively, as shown in Fig. 3B. To verify the predicted amino acid specificity, we overproduced and purified the NRPS-1A and NRPS-6A proteins (Fig. 3C) and examined their substrate specificity according to the amino acid-dependent ATP-PPi assay (Lee *et al.* (1970) *Meth. Enzymol.*, 43: 585-602; Ku *et al.* (1997) *Chem. & Biol.*, 4: 203-207). NRPS-1A and NRPS-6A indeed activate specifically L-Cys and L-Thr, respectively, among the amino acids tested (Fig. 3D). The latter results greatly enhanced our

confidence in predicting the substrate specificity of a NRPS module by the above method. We subsequently determined the substrate specificity for all the NRPS modules identified from the *blm* gene cluster and they in fact accounted for all nine amino acids required for BLM biosynthesis (Fig. 2).

Using the substrate specificity of individual NRPS and PKS modules as a guide, we can align the nine NRPS and one PKS modules to constitute the Blm megasynthetase as shown in Fig. 1B according to our hybrid NRPS/PKS/NRPS model for BLM biosynthesis (Fig. 1A). Among all the PKSs or NRPS systems examined so far, the Blm megasynthetase consists of the largest number of individual proteins. The precise interactions among all the Blm NRPS and Blm PKS proteins to constitute the Blm megasynthetase complex, therefore, reflect a remarkable power of protein-protein recognition (Guenzi *et al.* (1998) *J. Biol. Chem.* 273: 14403-14410; Gokhale *et al.* (1999) *Science* 284: 482-485). Although we are yet to provide direct evidence supporting the specific protein-protein interactions between the neighboring proteins, it is striking to note that all the biosynthetic intermediates isolated are derailed from either PKS or NRPS modules at the junctions between the interacting proteins (Fig. 1B). Since it is not difficult to imagine that an intermediate is more likely to fall off the enzyme complex when it is subjected to interpeptide transfer than to intrapeptide transfer, we view the latter observation as strong evidence supporting the current model of the Blm megasynthetase

BlmIX/BlmVIII/BlmVII as a hybrid NRPS/PKS/NRPS model.

Recent biosynthetic studies on rapamycin in *Streptomyces hygroscopicus* (Konig *et al.* (1997) *Eur. J. Biochem.* 247: 526-534), yersiniabactin in *Yersinia enterocolitica* and *Y. pestis* (Pelludat *et al.* (1998) *J. Bacteriol.* 180: 538-546; Gehring *et al.* (1998) *Chem. Biol.* 5: 573-586; Gehring *et al.* (1998) *Biochemistry* 37: 11637-11650) and TA in *Myxococcus xanthus* (Paitan *et al.* (1999) *J. Mol. Biol.* 286, 465-474) are starting to shed light on hybrid peptide and polyketide biosynthesis. Two models are emerging for the alignment between a NRPS and a PKS module. The interacting NRPS and PKS modules could be either covalently linked by arranging all domains in a linear order on the same protein (Pelludat *et al.* (1998) *J. Bacteriol.* 180: 538-546; Gehring *et al.* (1998) *Chem. Biol.* 5: 573-586; Gehring *et al.* (1998) *Biochemistry* 37: 11637-11650; Paitan *et al.* (1999) *J. Mol. Biol.* 286: 465-474) or physically located on two separate proteins, requiring specific protein-protein recognition to ensure the correct pairing between the interacting modules (Pelludat *et al.* (1998) *J. Bacteriol.* 180: 538-546; König *et al.* (1997) *Eur. J. Biochem.* 247: 526-534;

Gehring *et al.* (1998) *Chem. Biol.* 5: 573-586; Gehring *et al.* (1998) *Biochemistry* 37: 11637-11650). Common to all these systems, however, are the unusual features associated with the interacting modules, such as the lack of the AT domain of the PKS module in TaI (Paitan *et al.* (1999) *J. Mol. Biol.* 286: 465-474) and the lack of the A domain and the presence of the Cy domain of the NRPS modules in both HMWP1 and HMWP2 (Pelludat *et al.* (1998) *J. Bacteriol.* 180: 538-546; Gehring *et al.* (1998) *Chem. Biol.* 5: 573-586; Gehring *et al.* (1998) *Biochemistry* 37: 11637-11650). While extremely intriguing, the latter features complicate mechanistic analysis of these systems, making them less ideal candidates for studying how NRPS and PKS integrate into a productive hybrid NRPS/PKS complex.

The *BlmIX/BlmVIII/BlmVII* system combines the features of both hybrid NRPS/PKS and PKS/NRPS systems, serving as an ideal model for studying hybrid peptide and polyketide biosynthesis. The fact that both the *BlmIX* and *BlmVII* NRPS modules and the *BlmVIII* PKS module themselves are three separate proteins with a typical domain organization for NRPS and PKS enzymes greatly simplifies the mechanistic analysis of the hybrid NRPS/PKS/NRPS complex. We have found that the KS domain of *BlmVIII* is more similar to the KSs of HMWP1 (Pelludat *et al.* (1998) *J. Bacteriol.* 180: 538-546) and TaI (Paitan *et al.* (1999) *J. Mol. Biol.* 286: 465-474), both of which catalyze the elongation of a peptidyl intermediate with a malonate, than to KSs of type I PKSs. We attribute these subtle differences to their unique reactivity that catalyzes the transfer of the peptidyl intermediate from the PCP to the KS domain, which presumably takes place prior to chain elongation (Fig.4). Subsequent condensation catalyzed by the KS domain between the peptidyl intermediate and malonyl-S-ACP results in the elongation of the growing peptide with a carboxylic acid. Equally striking are the discoveries that the ACP domain of *BlmVIII* is more similar to a PCP than to an ACP and that the C domain of *BlmVII* has an additional N-terminal segment of about 50 amino acids that is rich in arginine, aspartic acid, and glutamic acid. The latter feature is analogous to the N-terminal interpolypeptide linker for type I PKS, which has recently been demonstrated to play a critical role in intermodular communication (Gokhale *et al.* (1999) *Science* 284: 482-485). We propose that these unique features of the ACP domain from the *BlmVIII* PKS module and the C domain from the *BlmVII* NRPS module provide the molecular basis for the C domain to recognize the acyl-S-ACP as a substrate. Subsequent condensation catalyzed by the C domain between acyl-S-ACP and amino acyl-S-PCP results in the elongation of the growing polyketide (as far as this condensation is concerned) with an amino acid (Fig. 4).

Novel domains for the Blm NRPS and PKS modules.

Various NRPS and PKS domains have been characterized, which are the building blocks for the entire field of combinatorial biosynthesis. The success for combinatorial biosynthesis depends critically upon the repertoire of these individual domains. Genetic analysis of the *blm* gene cluster has uncovered several novel NRPS and PKS domains. Without being bound to a particular theory, it is believed that *BlmVI* and *BlmV* are involved in the biosynthesis of the β -aminoalaninamide and pyrimidine moieties of BLM). In addition, the MT domain in *BlmVIII*, the Cy domains in *BlmIV*, and the Ox domain in *BlmIII* are novel domains.

The *BlmVIII* PKS module apparently furnishes the “propionate” unit into BLM in two steps by evolving a malonyl CoA-specifying AT domain coupled with a novel S-adenosylmethionine-requiring MT domain, representing a new mechanism to introduce methyl branches into polyketides (Fig. 4). This biosynthetic reaction sequence is unprecedented for polyketide biosynthesis since all PKSs from actinomycetes examined to date incorporate the alkyl branches into the resultant polyketides by selecting various alkyl malonates as the extending units that are determined by the AT domains. Yet, feeding experiments have unambiguously established that the polyketide moiety of BLM was derived from an acetate and a methionine (Takita and Muroka (1990) pages 289-309 in *Biochemistry of Peptide Antibiotics: Recent Advances in the Biotechnology of β -Lactams and Microbial Peptides*, Kleinkauf, H. & von Döhren, H. eds., W. de Gruyter, N.Y.), a fact that fits well with the observed unusual domain organization of the *BlmVIII* PKS module (Fig. 4). It is conceivable that the combination of this MT domain with an AT domain specific for a methyl malonate extending unit (Haydock *et al.* (1995) *FEBS Lett.* 374: 246-248) could result in the synthesis of polyketides with a gem-dimethyl moiety via engineering polyketide biosynthesis. Such a gem-dimethyl group has been found to be a very important pharmacophore for the epothilones, a family of hybrid peptide and polyketide metabolites that exhibits a remarkable antitumor activity similar to taxol (Ojima *et al.* (1999) *Proc. Natl. Acad. Sci. USA* 96: 4256-4261).

The *BlmIV* and *BlmIII* NRPSs are characterized by the unusual Cy domains as well as the unprecedented Ox domain, providing an efficient biosynthesis for a bithiazole structure. The Cy domain was first defined by Marahiel and co-workers in their study of bacitracin biosynthesis in *B. licheniformis* (Konz *et al.* (1997) *Chem. Biol.* 4: 927-937), and the Cy activity was demonstrated recently by Walsh and co-workers in their study of the

HMWP1 and HMWP2 proteins for yersiniabactin biosynthesis in *Y. pestis* (Gehring *et al.* (1998) *Chem. Biol.* 5: 573-586; Gehring *et al.* (1998) *Biochemistry* 37: 11637-11650).

While thiazoline is the direct product of the Cy domain, the thiazoline-to-thiazole conversion requires an additional oxidation step. We identified at the C-terminus of NRPS-0 an

5 additional domain that shows low, but significant, sequence homology to a family of putative oxidases/dehydrogenases, including the MbcC protein of the microcin B17 synthase (Table 1). Microcin B17 synthase catalyzes the synthesis of the oxazole and thiazole-containing peptide antibiotic microcin B17, and MbcC has been proposed to play a role in catalyzing the oxazoline/thiazoline-to-oxazole/thiazole conversion (Li *et al.* (1996) *Science* 274: 1188-

10 1193; Milne, *et al.* (1999) *Biochemistry* 38: 4768-4781). Consequently, we propose that this extra domain at the C-terminus of NRPS-0 could provide the oxidase/dehydrogenase activity needed for the biosynthesis of the bithiazole moiety of BLM, defining a novel Ox domain for NRPSs. It is noteworthy that a cell-free preparation from *Sv* ATCC15003 has been reported to catalyze the conversion of phleomycins to BLMs in the presence of NAD⁺ (Takita and

15 Muroka (1990) pages 289-309 in *Biochemistry of Peptide Antibiotics: Recent Advances in the Biotechnology of β -Lactams and Microbial Peptides*, Kleinkauf, H. & von Döhren, H. eds., W. de Gruyter, N.Y.), supporting the hypothesis that the bithiazole moiety of BLM results from stepwise oxidations of a bithiazoline precursor (Fig. 1A). (The phleomycin producer could be imagined to result from the loss of its Ox activity for the first thiazoline

20 ring.) Given the wide distribution of thiazole or oxazole rings in natural products (Ojima *et al.* (1999) *Proc. Natl. Acad. Sci. USA* 96: 4256-4261; Li *et al.* (1996) *Science* 274: 1188-1193) exhibiting an impressive array of biological activities, the cloning of the *blmIV,III* genes and the identification of the Ox domain open many opportunities to define the mechanism for thiazole biosynthesis and to potentially synthesize novel thiazole containing

25 molecules by engineering peptide biosynthesis.

Example 2

Identification and characterization of a type II peptidyl carrier protein from the bleomycin producer *Streptomyces verticillus* ATCC 15003.

Results.

Cloning and sequence analysis of the *blmI* gene

In our effort to clone the gene cluster responsible for BLM biosynthesis, we have determined 80 kb DNA sequence from *Sv* ATCC15003 (Fig. 8). Among the orfs identified within the *blm* gene cluster is the small orf of 273 base pairs (bp), *blmI*, which is located approximately 4 kb upstream of the previously characterized *blmAB* resistance locus (Sugiyama *et al.* (1994) *Gene* 151: 11-16; Calcutt and Schmidt (1994) *Gene* 151: 17-21) (Fig. 8B). The *blmI* gene encodes a protein of 90 amino acids with a molecular weight of 9957 and a pI of 6.52 (Fig. 8C). Computer-assisted analysis (Altschul *et al.* (1997) *Nucleic Acids Res.* 25: 3389-3402) of the deduced amino acid sequence indicates that BlmI is very similar to various PCP domains of NRPSs (ranging around 40% identity and 60% similarity, as shown in Figure 9). Like known PCP domains of NRPS, BlmI has the highly conserved signature motif of LGGXS, within which the serine residue is the site for 4'-phosphopantetheinylation (Stachelhaus and Marahiel (1995) *FEMS Microbiol. Lett.* 125: 3-14; Marahiel *et al.* (1997) *Chem. Rev.* 97: 2651-2673). The latter posttranslational modification is generally necessary for peptide biosynthesis; converting the apo-PCP into the functional holo-PCP (Marahiel *et al.* (1997) *Chem. Rev.* 97: 2651-2673; Walsh *et al.* (1997) *Curr. Opin. Chem. Biol.* 1: 309-315). Based on sequence comparison, *BlmI* is most related to PCPs and not to other kinds of carrier proteins that also share the same LGGXS (SEQ ID NO:80) motif and undergo the same posttranslational 4'-phosphopantetheinylation [31], such as the *E. coli* acyl carrier protein (ACP) (Lambalot and Walsh (1995) *J. Biol. Chem.* 270: 24658-24661), the ACP domain of type I PKS and the type II PKS ACP (Cox and Simpson (1997) *FEBS Lett.* 405: 267-272; Carreras *et al.* (1997) *Biochemistry* 36: 11757-11761), the ArCP domain (Gehring *et al.* (1998) *Biochemistry* 37: 2648-2659), and several nodulation related ACP-like proteins (Epple *et al.* (1998) *J. Bacteriol.* 180: 4950-4954; Spaink *et al.* (1991) *Nature* 354: 125-130).

Overexpression of *blmI* in *E. coli*

To overexpress the *blmI* gene in *E. coli*, we directly amplified the *blmI* gene by PCR from the *Sv. ATCC15003* genomic DNA and cloned it into the pQE-60 expression vector to give pBS1 so that BlmI could be produced as a protein with a native N-terminus and a His₆-tag at its C-terminus. However, no production of the BlmI protein was detected, as judged by sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE), upon introduction of pBS1 into *E. coli* M15(pREP4) under the standard overexpression conditions recommended by the manufacturer (Qiagen). We reasoned that the small BlmI protein with its native N-terminus may not be stable in the heterologous host, and hence moved the *blmI* gene from pBS1 into pET-29a to yield the second overexpression construct of pBS2. In the latter construct, BlmI should be produced as a fusion protein with 27 extra amino acid residues at its N-terminus, including an S-tag and the thrombin cleaving site, in addition to the His₆-tag at its C-terminus. Introduction of pBS2 into *E. coli* BL21(DE-3) under the standard overexpression conditions recommended by the manufacturer (Novagen) indeed resulted in overproduction of BlmI. In fact, the bulk of the soluble protein was the overproduced BlmI, which was easily purified by affinity chromatography using Ni-NTA resin (Qiagen). It is noteworthy that fusion of the additional 23 amino acids to the N-terminus of BlmI as in pBS2 and change of the expression system from *E. coli* M15(pREP4) (pBS1) to *E. coli* BL21(DE-3)(pBS2) dramatically improved the expression level of *blmI*.

In vivo 4'-phosphopantetheinylation of the BlmI protein

To establish BlmI as a type II PCP, we tested if it could serve as a substrate for a PCP-specific 4'-PPTase. PPTases catalyze the posttranslational modification of an apo-PCP into a holo-PCP by transferring the 4'-phosphopantetheine moiety from co-enzyme A (CoA) to the conserved serine residue of PCP, and this reaction has been developed recently into a general method to prepare various holo-PCP, holo-ACP, or holo-ArCP from the corresponding apoproteins (Stachelhaus *et al.* (1996) *Chem. Biol.* 3: 913-921; Gehring *et al.* (1998) *Biochemistry* 37: 2648-2659; Gehring *et al.* (1998) *Biochemistry* 37: 11637-11650; Weinreb *et al.* (1998) *Biochemistry* 37: 1575-1584). Therefore, we decided to investigate the 4'-phosphopantetheinylation of BlmI under both *in vivo* (Ku *et al.* (1997) *Chem. Biol.* 4: 203-207) and *in vitro* (Gehring *et al.* (1998) *Biochemistry* 37: 11637-11650; Lambalot *et al.* (1996) *Chem. Biol.* 3: 923-936; Quadri *et al.* (1998) *Biochemistry* 37: 1585-1595) conditions.

To examine 4'-phosphopantetheinylation of BlmI *in vivo*, we chose *E. coli* OG7001 as the expression host, which is a β -alanine auxotroph derived from *E. coli* BL21(DE3) by P1 co-transduction of the *panD* mutation from *E. coli* SJ16 (Epple *et al.* (1998) *J. Bacteriol.* 180: 4950-4954). Upon introduction of pBS2 into *E. coli* OG7001, *blmI* was exceptionally well expressed and the overproduced BlmI protein was readily purified. However, high performance liquid chromatography (HPLC) analysis showed that the purified BlmI was essentially in the apo-form (Fig. 10A), indicative that apo-BlmI was a poor substrate for the *E. coli* endogenous PPTases, such as EntD and ACP synthase (Lambalot *et al.* (1996) *Chem. Biol.* 3: 923-936; Walsh *et al.* (1997) *Curr. Opin. Chem. Biol.* 1: 309-315; Lambalot and Walsh (1995) *J. Biol. Chem.* 270: 24658-24661). To circumvent the poor endogenous PPTase activity, we next co-expressed *blmI* with the *gsp* gene, which was isolated from the gramicidin S producer *Bacillus brevis*, and encoded a PPTase that was known to 4'-phosphopantetheinylate heterologously produced PCPs in *E. coli* (Lambalot *et al.* (1996) *Chem. Biol.* 3: 923-936; Ku *et al.* (1997) *Chem. Biol.* 4: 203-207). We co-transformed pDPT-Gsp, in which the expression of the *gsp* gene was under the control of the T5/Lac promoter (Ku *et al.* (1997) *Chem. Biol.* 4: 203-207), and pBS2 into *E. coli* OG7001. BlmI was again very well expressed and the resulting BlmI protein was similarly purified. HPLC analysis showed that at least 60% of overproduced BlmI was modified into the holo-BlmI protein (Fig. 10B). (A PCP domain was similarly 4'-phosphopantetheinylated *in vivo* before by co-expressing *gsp* in *E. coli* using pDPT-Gsp, and approximately 80% of the PCP was produced in the holo-form (Ku *et al.* (1997) *Chem. Biol.* 4: 203-207).

We next cultured *E. coli* OG7001(pBS2) and *E. coli* OG7001(pBS2/pDPT-Gsp) in the presence of [3-³H]- β -alanine, a known biosynthetic precursor of 4'-phosphopantetheine (Stachelhaus *et al.* (1996) *Chem. Biol.* 3: 913-921; Epple *et al.* (1998) *J. Bacteriol.* 180: 4950-4954). Specific incorporation of [3-³H]- β -alanine into the 4'-phosphopantetheine moiety of holo-BlmI was determined by autoradiographic analysis. Thus, while fermentation of *E. coli* OG7001(pBS2) in the presence of [3-³H]- β -alanine led to an IPTG-dependent overproduction of BlmI, little of the resulting BlmI protein was ³H-labeled, indicative of being produced in the apo-form. In contrast, fermentation of *E. coli* OG7001(pBS2/pDPT-Gsp) in the presence of [3-³H]- β -alanine resulted in a significant increase of IPTG-dependent incorporation of the ³H-label into the overproduced BlmI protein, suggesting a specific incorporation of [3-³H]- β -alanine into holo-BlmI, presumably in the 4'-phosphopantetheine moiety. There were several additional proteins that were also

weakly labeled by $[3\text{-}^3\text{H}]\text{-}\beta\text{-alanine}$. However, both their expression and their incorporation by ^3H -label were independent from either IPTG induction or the presence of Gsp, hence these proteins were unrelated to BlmI. (Similar background labeling was reported before for in vivo 4'-phosphopantetheinylation of other PCP (Epple *et al.* (1998) *J. Bacteriol.* 180: 4950-4954)). We also purified the BlmI protein from *E. coli* OG7001(pBS2/pDPT-Gsp) and demonstrated that it was the holo-BlmI protein that was specifically associated with the ^3H -activity. Finally, we confirmed the identity of holo-BlmI by subjecting the purified BlmI protein to MALDI-Tof mass spectral analysis (Weinreb *et al.* (1998) *Biochemistry* 37: 1575-1584). BlmI produced in the absence of the Gsp PPTase yielded a single peak with a molecular weight of 13,952, suggesting that the produced BlmI protein is in the apo-form (calc., 13,949). In contrast, BlmI produced in the presence of Gsp yielded two species with molecular weight of 13,969 and 14,303, respectively. While the species with the molecular weight of 13,969 represents apo-BlmI, a molecular weight of 14,303 unambiguously confirmed the other protein as holo-BlmI (calc., 14,289). The latter result indicated that the purified BlmI consisted of both the apo- and holo-BlmI proteins, in agreement with the HPLC analysis results (Fig. 10B).

In vitro 4'-phosphopantetheinylation of the BlmI protein

To investigate 4'-phosphopantetheinylation of BlmI in vitro, we chose the Sfp protein as the preferred PPTase, which had been isolated before from the surfactin producer *Bacillus subtilis* (Nakano *et al.* (1992) *Mol. Gen. Genet.* 232: 313-321). (Overexpression of *gsp* in *E. coli* using pDPT-Gsp resulted in predominantly an insoluble Gsp protein (Ku *et al.* (1997) *Chem. Biol.* 4: 203-207). The Sfp PPTase was overproduced in *E. coli* MV1190(pUC8-Sfp) and purified to near homogeneity as described before (Quadri *et al.* (1998) *Biochem.*, 37: 1585-1595; Nakano *et al.* (1992) *Mol. Gen. Genet.*, 232: 313-321). Upon incubation of the purified apo-BlmI with $[^3\text{H}\text{-panthetheine}]\text{-CoA}$ in the presence of the Sfp PPTase, we examined the covalent incorporation of the $[^3\text{H}\text{-panthetheine}]\text{-4'}$ -phosphopantetheine moiety from CoA into holo-BlmI by autoradiographic analysis. Indeed, the apo-BlmI was quantitatively labeled by $[^3\text{H}\text{-panthetheine}]\text{-CoA}$, and no labeling was observed in the absence of either the apo-BlmI or the Sfp PPTase protein, demonstrating that the Sfp PPTase can recognize apo-BlmI as a substrate and specifically transfer the 4'-phosphopantetheine group from CoA into holo-BlmI.

In vitro aminoacylation of BlmI

Once we established BlmI as a type II PCP that can be readily modified by PCP-specific PPTases into the holo-BlmI protein, we tested if the holo-BlmI could be aminoacylated in trans, requiring an A domain. Since BlmI has no cognate A domain of its own, we turned our attention to another putative biosynthesis gene cluster we have cloned previously from *Sv* ATCC15003, which encodes at least four NRPS and one PKS modules. We have established that this gene cluster is not clustered with the *blm* locus and is unrelated to BLM biosynthesis. From this gene cluster, we amplified by PCR a 1579 bp fragment encoding an A domain, named Val-A, which we predicted to have a molecular weight of 56,581 and a pI of 7.39. We cloned *val-A* into pET-28a to yield pBS3, in which Val-A would be produced as a fusion protein with a His₆-tag at the N-terminus. Introduction of pBS3 into *E. coli* BL21(DE3) under the standard overexpression conditions recommended by the manufacturer (Novagen) resulted in good overproduction of Val-A, predominantly in soluble form, from which Val-A was purified by affinity chromatography using Ni-NTA resin. The purified Val-A protein was active by the amino acid-dependent ATP-PPi exchange assay (Lee and Lipmann (1970) *Method Enzymol.* 43: 585-602; Ku *et al.* (1997) *Chem. Biol.*, 4: 203-207). Among the 23 amino acids tested, Val-A specifically activated valine, an amino acid that is not required for BLM biosynthesis.

To carry out the aminoacylation in trans, we incubated the purified holo-BlmI and Val-A in vitro in the presence *L*-[¹⁴C(U)]valine and ATP (Stachelhaus *et al.* (1996) *Chem. Biol.* 3: 913-921; Weinreb *et al.* (1998) *Biochemistry* 37: 1575-1584). The aminoacylated holo-BlmI-*L*-[¹⁴C(U)]valine species was subjected to SDS-PAGE and specific attachment of *L*-[¹⁴C(U)]valine to holo-BlmI was determined by autoradiographic analysis. Remarkably, the holo-BlmI was specifically labeled by *L*-[¹⁴C(U)]valine in the presence of Val-A, indicative of the formation of the holo-BlmI-S-valine thioester. The in trans aminoacylation between the holo-BlmI and Val-A proteins appeared to be very specific. Neither incubation of *L*-[¹⁴C(U)]valine with Val-A, the apo-BlmI, or the holo-BlmI protein alone, nor incubation of *L*-[¹⁴C(U)]valine with the Val-A and apo-BlmI proteins, resulted in the detection of ¹⁴C-labeled BlmI protein.

Discussion.

Nonribosomal peptides and polyketides are two distinct classes of natural products yet are assembled from amino acids and short carboxylic acids by NRPSs and PKSs, respectively, in strikingly similar strategies (Cane *et al.* (1998) *Science* 282: 63-68).

These fascinating multifunctional enzyme complexes have been classified into two types based on their gene organization and enzyme architecture. Type I enzymes are multifunctional proteins consisting of domains for individual enzyme activities, and type II enzymes are multienzyme complexes consisting of discrete proteins that are largely monofunctional. While both type I and type II PKSs (Fig. 11A and 11C) have been well characterized to account for the vast structural diversities found in polyketide biosynthesis (Hopwood (1997) *Chem. Rev.* 97: 2465-2497), all NRPSs studied so far are exclusively the type I modular enzymes (Fig. 11B) (Kleinkauf and von Döhren: H. (1996) *Eur. J. Biochem.* 236: 335-351; Marahiel *et al.* (1997) *Chem. Rev.* 97: 2651-2673; von Döhren *et al.* (1997) *Chem. Rev.* 97: 2675-2705). It is very tempting to speculate the existence of a type II NRPS that, analogous to type II PKS (Shen and Hutchinson (1993) *Science* 262: 1535-1540; Bao *et al.* (1998) *Biochemistry* 37: 8132-8138; Carreras and Khosla (1998) *Biochemistry* 37: 2084-2088), should consist of discrete proteins possessing enzyme activities such as the A (Stachelhaus and Marahiel (1995) *J. Biol. Chem.* 270: 6163-6169), the PCP (Stein and Morris (1996) *J. Biol. Chem.* 271: 15428-15435), or the C (Stachelhaus *et al.* (1998) *J. Biol. Chem.* 273: 22773-22781) domains of type I NRPSs (Fig. 11D). The fact that both the A (Stachelhaus and Marahiel (1995) *J. Biol. Chem.* 270: 6163-6169; Konz *et al.* (1997) *Chem. Biol.* 4: 927-937; Weinreb *et al.* (1998) *Biochemistry* 37: 1575-1584; Mootz and Marahiel (1997) *J. Bacteriol.* 179: 6843-6850) and the PCP (Stachelhaus *et al.* (1996) *Chem. Biol.* 3: 913-921; Weinreb *et al.* (1998) *Biochemistry* 37: 1575-1584; Pfeifer *et al.* (1995) *Biochemistry* 34: 7450-7459; Haese *et al.* (1994) *J. Mol. Biol.* 243: 116-122; Lambalot *et al.* (1996) *Chem. Biol.* 3: 923-936; Quadri *et al.* (1998) *Biochemistry* 37: 1585-1595; Gehring *et al.* (1996) *Chem. Biol.* 4: 17-24; Ku *et al.* (1997) *Chem. Biol.* 4: 203-207) domains of type I NRPSs can act as independent enzymes supports the hypothesis of a type II NRPS.

We have now cloned and sequenced the *blmI* gene, overproduced and characterized the BlmI protein as a bona fide type II PCP, and demonstrated that holo-BlmI can be aminoacylated by a completely unrelated A domain, providing for the first time genetic and biochemical evidence for a type II NRPS enzyme. We concluded BlmI as a type II PCP based on the following criteria. (1) The deduced amino acid sequence of the *blmI* gene is highly homologous to various PCP domains of known NRPSs, in particular at the signature motif of LGGXS within which the 4'-phosphopantetheine prosthetic group is covalently attached to the serine residue (Marahiel *et al.* (1997) *Chem. Rev.* 97: 2651-2673; Stachelhaus and Marahiel (1995) *FEMS Microbiol. Lett.* 125: 3-14). While the current boundaries for a PCP domain in the literature were defined arbitrarily (Stachelhaus *et al.*

(1996) *Chem. Biol.* 3: 913-921) and varied from one PCP to another, we can now re-define a PCP domain for the type I NRPS as a 90 amino acid peptide with approximately 45 amino acids, each flanking the essential serine residue in the LGGXS (SEQ ID NO:81) motif, in light of this discrete BlmI type II PCP (Fig.9). (2) The *blmI* gene has been successfully expressed in *E. coli*, and fusion of a short peptide to the N-terminus of BlmI dramatically improved its overproduction efficiency. While we cannot exclude the effect of different systems on gene expression, i.e., *E. coli* M15(pREP4)(pBS1) vs. *E. coli* BL21(DE-3)(pBS2), we attribute the increase in expression efficiency to the stability of BlmI as an N-terminal fusion protein instead of the otherwise labile BlmI protein with its native N-terminus. Since BlmI was produced predominantly in the apo-form in *E. coli*, apo-BlmI apparently was not a substrate for the endogenous PPTases, such as EntD or ACP synthase, excluding BlmI as an ArCP or ACP, respectively. EntD and ACP synthase are known to 4'-phosphopantetheinylate apo-ArCP and ACP, respectively, to their holo-forms efficiently (Lambalot *et al.* (1996) *Chem. Biol.* 3: 923-936; Walsh *et al.* (1997) *Curr. Opin. Chem. Biol.* 1: 309-315; Lambalot and Walsh (1995) *J. Biol. Chem.* 270: 24658-24661). (3) The apo-BlmI protein serves as a substrate for PCP-specific PPTases that transfer the 4'-phosphopantetheine moiety from CoA to apo-BlmI to yield the holo-BlmI protein. We have demonstrated this posttranslational modification for BlmI *in vivo* with the Gsp PPTase (Ku *et al.* (1997) *Chem. Biol.* 4: 203-207) and *in vitro* with the Sfp PPTase (Gehring *et al.* (1998) *Biochemistry* 37: 11637-11650; Lambalot *et al.* (1996) *Chem. Biol.* 3: 923-936; Quadri *et al.* (1998) *Biochemistry* 37: 1585-1595), both of which have been extensively used in preparing holo-PCPs. (4) The specific modification of apo-BlmI by 4'-phosphopantetheinylation has been monitored by HPLC analysis (Fig. 10) (Weinreb *et al.* (1998) *Biochemistry* 37: 1575-1584) and by specific incorporation of [3-³H]- β -alanine *in vivo* (Stachelhaus *et al.* (1996) *Chem. Biol.* 3: 913-921; Ku *et al.* (1997) *Chem. Biol.* 4: 203-207; Eppele *et al.* (1998) *J. Bacteriol.* 180: 4950-4954) and of [³H-pantetheine]-CoA *in vitro* (Gehring *et al.* (1998) *Biochemistry* 37: 11637-11650; Lambalot *et al.* (1996) *Chem. Biol.* 3: 923-936; Quadri *et al.* (1998) *Biochemistry* 37: 1585-1595), respectively, into the 4'-phosphopantetheine moiety of the holo-BlmI protein. The identity of BlmI was finally confirmed by MALDI-ToF mass spectral analysis that determined the molecular weight for both the apo- and holo-BlmI proteins.

While individual domains of type I NRPSs can function independently and several A (Stachelhaus and Marahiel (1995) *J. Biol. Chem.* 270: 6163-6169; Konz *et al.*

(1997) *Chem. Biol.* 4: 927-937; Weinreb *et al.* (1998) *Biochemistry* 37: 1575-1584; Mootz and Marahiel (1997) *J. Bacteriol.* 179: 6843-6850) and PCP (Stachelhaus *et al.* (1996)

Chem. Biol. 3: 913-921; Weinreb *et al.* (1998) *Biochemistry* 37: 1575-1584; Pfeifer *et al.* (1995) *Biochemistry* 34: 7450-7459; Haese *et al.* (1994) *J. Mol. Biol.* 243: 116-122;

5 Lambalot *et al.* (1996) *Chem. Biol.* 3: 923-936; Quadri *et al.* (1998) *Biochemistry* 37: 1585-1595; Gehring *et al.* (1996) *Chem. Biol.* 4: 17-24; Ku *et al.* (1997) *Chem. Biol.* 4: 203-207) domains have been overproduced, purified, and biochemically characterized, aminoacylation in trans has been successful only between PCPs and their cognate A domains (Stachelhaus *et al.* (1996) *Chem. Biol.* 3: 913-921; Weinreb *et al.* (1998) *Biochemistry* 37: 1575-1584). No

10 aminoacylation between PCP and A domains from different NRPS modules has been observed. These results led to the conclusion that there is a specific protein-protein recognition between the A domain and its cognate PCP (Weinreb *et al.* (1998) *Biochemistry* 37: 1575-1584). Such domain-specific aminoacylation, in fact, should be beneficial in maintaining the fidelity of a type I NRPS by providing additional "gating" against

15 misincorporation of non-specifically activated aminoacyl adenylate into the final peptide product. Since a type II PCP such as BlmI lacks its cognate A domain, we asked if BlmI could be aminoacylated by an unrelated A domain of a type I NRPS. Although we have yet to determine the biochemical role of BlmI in vivo, the fact that the *blmI* gene is located in the middle of the *blm* gene cluster suggests that it may be involved in BLM biosynthesis. To

20 avoid the ambiguity of selecting an A domain that may potentially interact with BlmI in vivo, we preferred not to choose any A domain from the *blm* gene cluster to test if it could aminoacylate BlmI in trans. We reasoned that an A domain that is unrelated to BlmI should come from a gene cluster independent from BLM biosynthesis and should activate an amino acid not required by BLM. We chose Val-A because it satisfied both requirements. Val-A is

25 an A domain of a type I NRPS from a gene cluster we have cloned previously from *Sv* ATCC15003 that has proven to be unrelated to BLM biosynthesis, and it specifically activates valine among the 23 amino acids tested. Remarkably, BlmI was efficiently aminoacylated by Val-A. The valine residue is specifically attached in a thioester linkage to the terminal -SH of the 4'-phosphopantetheine moiety of the holo-BlmI protein, as evidenced

30 by the fact that the apo-BlmI was inactive under the identical conditions.

Aminoacylation of holo-BlmI by Val-A represents the first example in which an A domain aminoacylates a protein other than its cognate PCP domain. Since it has been suggested that an A domain of a type I NRPS can transfer the activated aminoacyl adenylate only to its cognate PCP domain because of the specific protein-protein recognition between

the two domains (Weinreb *et al.* (1998) *Biochemistry* 37: 1575-1584), the fact that BlmI is aminoacylated by Val-A revealed a distinct feature of a type II PCP. It is very tempting to speculate that type II PCPs such as BlmI may have broad intrinsic substrate specificity toward either the aminoacyl adenylate, the A domain, or both. In fact, the latter feature is reminiscent of the type II PKS ACPs, which have been shown to be interchangeable among different PKS complexes (Shen and Hutchinson (1993) *Science* 262: 1535-1540; Bao *et al.* (1998) *Biochemistry* 37: 8132-8138; Carreras and Khosla (1998) *Biochemistry* 37: 2084-2088). The biosynthesis of *D*-alanyl-lipoteichoic acid in *Bacillus subtilis* (Perego *et al.* (1995) *J. Biol. Chem.* 270: 15598-15606) and *Lactobacillus casei* (Debabov *et al.* (1996) 178: 3869-3876) also involves a discrete ACP-like protein, the *D*-alanyl carrier protein, although the latter clearly is structurally and functionally different from PCPs.

The results strongly suggest the existence of a type II NRPS. In fact, we have already identified within the *blm* gene cluster two additional genes, *blmII* and *blmXI* (Fig. 1B), which encode type II C proteins based on sequence analysis (*see* Example 1).

15 **Significance.**

All NRPSs known to date are exclusively the type I modular enzymes that are multifunctional proteins consisting of domains, such as A (Stachelhaus and Marahiel (1995) *J. Biol. Chem.* 270: 6163-6169), PCP (Stachelhaus *et al.* (1996) *Chem. Biol.* 3: 913-921), and C (Stachelhaus *et al.* (1998) *J. Biol. Chem.* 273: 22773-22781), for individual enzyme activities (Kleinkauf and von Döhren: H. (1996) *Eur. J. Biochem.* 236: 335-351; Marahiel *et al.* (1997) *Chem. Rev.* 97: 2651-2673; von Döhren *et al.* (1997) *Chem. Rev.* 97: 2675-2705), and control the structural variations of the resulting peptide products by the multiple-carrier thiotemplate mechanism (Cane *et al.* (1998) *Science* 282: 63-68; Stein and Morris (1996) *J. Biol. Chem.* 271: 15428-15435). While individual domains of type I NRPSs can function independently, aminoacylation *in trans* has been successful only between PCPs and their cognate A domains (Stachelhaus *et al.* (1996) *Chem. Biol.* 3: 913-921; Weinreb *et al.* (1998) *Biochemistry* 37: 1575-1584). We have cloned and sequenced the *blmI* gene, overproduced and characterized the BlmI protein as a bona fide type II PCP, and demonstrated that the holo-BlmI can be aminoacylated by a completely unrelated A domain. Our results provided for the first time the genetic and biochemical evidence to support the hypothesis of a type II NRPS, setting the stage for formulating new research concepts to study peptide biosynthesis. Genetic manipulation of type I NRPS has already been successful in generating novel peptides (Stachelhaus *et al.* (1995) *Science* 269: 69-72). An unprecedented type II NRPS

should shed new light in engineering NRPS proteins, greatly increasing our ability to access peptides with even greater structural diversities.

Materials and methods

General DNA manipulations

5 Plasmids preparation and DNA extraction were carried out by using commercial kits (Qiagen, Santa Clarita, CA), and all other manipulations were carried out according to standard methods (Sambrook *et al.* (1989) *Molecular cloning: a laboratory manual*: (2nd ed): Cold Spring Harbor Laboratory Press: Cold Spring Harbor: USA). *E. coli* strain DH5 α was used as the host for general DNA propagations.

Overexpression of *blmI* in *E. coli* and purification of the BlmI protein

10 The *blmI* gene was amplified from Sv ATCC15003 by PCR using a forward primer of 5'-CCG CCC ATG GGT GCT CCG CGT GGC GAG CGG ACC CGG CGC-3' (SEQ ID NO:82, the *NcoI* site is underlined) and a reverse primer of 3'-CCT AGA TCT CCG GTC CCG CTC CCC CGT-5' (SEQ ID NO:83, the *BglII* site is underlined). In order
15 to create the *NcoI* site, the original starting sequence of "ATG AGC" has been changed to "ATG GGT", which resulted in the change of the second amino acid from serine to glycine. The first five codons of *blmI* were also optimized for overexpression in *E. coli*. The PCR-amplified 0.3 kb *NcoI*-*BglII* fragment was cloned into the similar sites of pQE-60 (Qiagen) to form pBS1. Digestion of pBS1 with *NcoI* and *HindIII* and cloning the resulting 0.3 kb
20 *NcoI*-*HindIII* fragment into the same sites of pET-29a (Novagen, Madison, WI) yielded pBS2.

Expressions of *blmI* in *E. coli* M15 (pREP4)(pBS1) and in *E. coli* BL-21(DE-
3)(pBS2) and purification of the resulting BlmI protein by affinity chromatography on Ni-
25 NTA resin were carried out under the standard conditions recommended by Qiagen and Novagen, respectively. The incubation temperature was lowered to 30 °C to improve the solubility. The purification of BlmI was monitored by SDS-PAGE on 15% gel. The final pure BlmI protein was desalted on PD-10 column (Sephadex G-25, Pharmacia Biotech, Piscataway, NJ) into 50 mM sodium phosphate buffer, pH 7.8, containing 200 mM NaCl, 10
30 mM MgCl₂, 2 mM dithiothreitol (DTT), 1 mM EDTA, 10% glycerol, and stored at - 80 °C for in vitro assays.

HPLC analysis and MALDI-ToF mass spectral determination

Samples of BlmI (30–70 µg) purified from *E. coli* OG7001(pBS2) or *E. coli* OG7001(pBS2/pDPT-Gsp) were analyzed on a Nova-Pak C18 column (5mm x 10, Waters, Milford, MA) using a Rainin DMAX HPLC unit. The column was developed by a linear gradient of 0–50% acetonitrile in 0.1% trifluoroacetic acid in 25 min, followed by additional 5 min at 50 % acetonitrile, with a flow rate of 0.6 ml/min and detection at 280 nm. MALDI-ToF mass spectral determination was performed on a Bruker Biflex III spectrometer at the Facility for Advanced Instrumentation of University of California, Davis.

In vivo labeling of BlmI with [3-³H]-β-alanine

The β-alanine auxotroph *E. coli* strain OG7001 (Epple *et al.* (1998) *J. Bacteriol.* 180: 4950–4954) was transformed with pBS2 and cultured under the same conditions as for *E. coli* BL21(DE3) (Novagen). For co-expression of *blmI* with *gsp*, pDPT-Gsp (Ku *et al.* (1997) *Chem. Biol.* 4: 203–207) was similarly transformed into *E. coli* OG7001(pBS2) and the transformants were cultured in 2xYT (Debabov *et al.* (1996) 178: 3869–3876) in the presence of kanamycin (25 µg/ml) and chloramphenicol (50 µg/ml). For in vivo labeling experiment, cells from 2 ml overnight culture of either *E. coli* OG7001(pBS2) or *E. coli* OG7001(pBS2/pDPT-Gsp) were harvested, washed with M9 minimal medium (Debabov *et al.* (1996) 178: 3869–3876), and re-suspended in 2 ml of M9 minimal medium. The latter were used as seed cultures (20 µl) to inoculate 1 ml M9 medium with kanamycin (25 µg/ml) or kanamycin (25 µg/ml) and chloramphenicol (50 µg/ml) for *E. coli* OG7001(pBS2) or *E. coli* OG7001(pBS2/pDPT-Gsp), respectively. The resulting culture was incubated at 30 °C, 250 rpm to OD_{600nm} 0.6 and to this was added 10 µCi of [3-³H]-β-alanine (50 Ci/mmol, American Radiolabeled Chemicals Inc., St. Louis, MO) with or without IPTG (1 mM). Total proteins were resolved by SDS-PAGE on 15% gels that were Coomassie blue-stained. To determine ³H-labeling of the overproduced holo-BlmI protein, gels were soaked in Amplifier (Amersham, Arlington Heights, IL) for 20 min, dried between two sheets of cellulose membrane (KOH Development Inc., Ann Arbor, MI), and visualized by autoradiography on X-ray films (Fuji Medical Systems, Stamford, CT).

In vitro labeling of BlmI with [³H-pantetheine]-CoA

Expression of *sfp* in *E. coli* MV1190(pUC8-Sfp), purification of the Sfp PPTase to homogeneity, and 4'-phosphopantetheinylation of apo-BlmI by Sfp in vitro were

carried out essentially according to literature procedures (Quadri *et al.* (1998) *Biochemistry* 37: 1585-1595; Nakano *et al.* (1992) *Mol. Gen. Genet.* 232: 313-321). A typical 100 μ l assay solution contained 26 μ M apo-B1mI, 2.9 μ M Sfp, 25 μ M [3 H-pantetheine]-CoA (0.9 μ Ci, 40 Ci/mM), 10 mM $MgCl_2$, and 5 mM DTT, in 75 mM MES/NaOAc buffer, pH 6.0.

After 30 min incubation at 37 $^{\circ}$ C, the assays were stopped by addition of 5 μ l of bovine serum albumin (0.2 mg/ml) and 0.9 ml of cold 10% (v/v) trichloroacetic acid (TCA). The precipitated proteins were collected by centrifugation at 14,000 rpm, 20 min, 4 $^{\circ}$ C (Eppendorf 5415C centrifuge), washed with 10% TCA three times, and resolved by SDS-PAGE on 15% gel. The 3 H-activity incorporated into holo-B1mI was similarly determined by autoradiography as described for in vivo labeling of holo-B1mI with [3 -H]- β -alanine.

Overexpression of val-A in E. coli and purification and assay of the Val-A protein

The *val-A* fragment was amplified from *Sv* ATCC15003 by PCR using a forward primer of 5'-GGA ATT CCA TAT GGG CAC CAC CGT CGC CGC G-3' (SEQ ID NO:84, the *Nde*I site is underlined), and a reverse primer of 3'-GGC AAG CTT GGG ACC GGG CGT GGA GCG C (SEQ ID NO:85, the *Hind*III site is underlined). The PCR-amplified 1.6 kb *Nde*I-*Hind*III fragment was cloned in the similar sites of pET-28a (Qiagen) to yield pBS3. Expression of *val-A* in *E. coli* BL-21(DE-3)(pBS3) and purification of the resulting Val-A protein by affinity chromatography on Ni-NTA resin were carried out under the standard conditions recommended by Novagen.

Amino acid-dependent ATP-PPi assays were performed essentially according to the literature procedures (Ku *et al.* (1997) *Chem. Biol.* 4: 203-207; Lee and Lipmann (1970) *Method Enzymol.* 43: 585-602). A typical 100 μ l assay solution contained 180 nM Val-A, 1 mM ATP, 0.1 mM PPi with 0.2 μ Ci of 32 P-PPi (11.75 Ci/mmol, NEN Life Science Products, Inc., Boston, MA), 1 mM $MgCl_2$, 0.1 mM EDTA, and 1 mM L-amino acid in 50 mM sodium phosphate buffer, pH 7.8. After 30 min incubation at 30 $^{\circ}$ C, the assays were stopped by addition of 0.9 ml of cold 1% (w/v) activated charcoal in 3% (v/v) perchloric acid. The precipitates were collected on glass fiber filters (2.4 cm, G-4, Fisher, Pittsburgh, PA), washed successively with 10 ml of 0.2 M sodium phosphate buffer, pH 8.0, 4 ml water, and 1 ml of ethanol, and dried in air. The filters were mixed with 7 ml of scintillation fluid (ScintiSafe Gel, Fisher) and counted on a Beckman LS-6800 scintillation counter to determine the radioactivity.

In vitro aminoacylation of holo-BlmI by Val-A

The aminoacylation of holo-BlmI was carried out essentially according to literature methods (Stachelhaus *et al.* (1996) *Chem. Biol.* 3: 913-921; Weinreb *et al.* (1998) *Biochemistry* 37: 1575-1584). A typical 100 μ l assay solution contained 180 nM Val-A, 1.5-
5 2.8 μ M apo- or holo-BlmI, 35 μ M L-[14 C(U)]-valine (283 mCi/mmol, NEN Life Science Products, Inc., Boston, MA), 5 mM ATP, 10 mM MgCl₂, and 5 mM DTT in 75 mM Tris-HCl buffer, pH 8.0. The reactions were started by the addition of ATP and, after incubation at 37 °C for 30 min, were stopped by addition of 0.9 ml of cold 7% (v/v) TCA. The precipitated proteins were collected by centrifugation at 14,000 rpm, 20 min, 4 °C
10 (Eppendorf 5415C centrifuge) and resolved by SDS-PAGE on a 15% gel. The radioactivity incorporated into the holo-BlmI-L-[14 C(U)]valine species was similarly determined by autoradiography as described for in vivo labeling of holo-BlmI with [3 -H]- β -alanine.

Example 3:

Cloning and characterization of a phosphopantetheinyl transferase from the 15 bleomycin-producing *Streptomyces verticillus* ATCC15003

Multienzymes complexes exist for acyl group activation and transfer reactions in the biogenesis of fatty acids, the polyketide family of natural products (*e.g.* erythromycin, tetracycline), and almost all non-ribosomal peptides (*e.g.* vancomycin, cyclosporin, penicillin). All of these complexes contain one or more small proteins, ~80-100 amino acids
20 long, either as separate subunits or as integrated domains, that function as carrier proteins for the growing acyl chain (acyl-, peptidyl-, and aryl- carrier proteins, abbreviated as ACP, PCP, and ArCP). They are converted from inactive apo-forms to functional holo-forms by the covalent attachment of the 4'-phosphopantetheine moiety of coenzyme A to a conserved serine residue of the carrier-protein substrate. This essential post-translational modification
25 is catalyzed by a family of enzymes known as phosphopantetheinyl transferases (PPTases) (Lambalot *et al.* *Chem. Biol.* (1996) 3:923-936; Walsh *et al.* *Curr. Opin. Chem. Biol.* (1997) 1:309-315).

Research in the field of polyketide and non-ribosomal peptide biosynthesis has been hampered by the inability to fully modify and thus convert to the active form some
30 polyketide synthases (PKS) and polypeptide synthetases (NRPS) when overproduced in heterologous hosts, presumably because the host PPTases are unable to effectively modify these overexpressed protein substrates. Our group is currently involved in the

characterization of the gene cluster responsible for the biosynthesis of the antitumor drug bleomycin in *Streptomyces verticillus* ATCC15003. As bleomycin synthetase is a hybrid NRPS/PKS enzyme, we decided to obtain a PPTase from the producing organism in order to use it *in vitro* or *in vivo* by coexpression with the synthetase genes to produce properly modified, active synthetases for our studies.

Results and Discussion

Cloning of the *pta* gene from *S. verticillus* ATCC15003.

- The similarities among PPTases from different organisms are reduced to two short motifs separated by 40-45 residues: (V/I)G(V/I)D, and (F/W)(S/C/T)XKE(A/S)hhK (Lambalot et al. *Chem. Biol.* (1996) 3:923-936; Walsh et al. *Curr. Opin. Chem. Biol.* (1997) 1:309-315). Our previous attempts to amplify PPTase sequences from *S. verticillus* chromosomal DNA using degenerate primers according to the two conserved motifs were unsuccessful (unpublished results), so we decided to narrow our target. PPTases have been classified in two groups, according to their specificity for the carrier-protein substrate:
- 15 PPTases involved in polyketide/fatty acid biosynthesis use acyl carrier proteins (ACPs) as substrate, while those for non-ribosomal peptide biosynthesis use peptidyl carrier proteins (PCPs) or aryl carrier proteins (ArCPs) (Walsh et al. *Curr. Opin. Chem. Biol.* (1997) 1:309-315). Several "NRPS-type" PPTase sequences were used to screen the databases to look for actinomycete homologues, and four proteins of unknown function were found:
 - 20 NshC from *Streptomyces actuosus* (Li et al. *Gene* (1990) 91:9-17), SC5A7. 23 from *S. coelicolor* (GenBank AL031107), an unnamed protein from *Streptomyces* sp. strain TH1 (Mori et al. *J. Bacteriol.* (1997) 179:5677-5683), and Rv2794c (later renamed PptT (Quadri et al. *Chem. Biol.* (1998) 5:631-645)) from *Mycobacterium tuberculosis* (GenBank AL008967). The alignment of the actinomycete sequences showed the two motifs conserved
 - 25 in all PPTases and an additional motif - the "THC" motif: PXWPXGX₂GS(M/L)THCXGY (SEQ ID NO:86), located about 15 amino acids upstream of the (V/I)G(V/I)D motif (SEQ ID NO:87). The "THC" motif is not universally conserved in all PPTases, but it can be detected also in some non-actinomycete PPTases like EntD (Coderre et al. *J. Gen. Microbiol.* (1989) 135:3043-3055). Using a recently developed method of PCR primer design (the
 - 30 CODEHOP strategy (Consensus-DEgenerate Hybrid Oligonucleotide Primer) (Rose et al. *Nucleic Acids Res.* (1998) 26:1628-1635), two primers were designed around the typical C-terminal PPTase motif (primers KEA-1: 5'-T GCA GCA GAA CAG GAG GCK NYC CCA

NKG-3' (SEQ ID NO:88) and KEA-2: 5'-TG GGT CAG CGG GTA CCA NRC YTT RWA-3' (SEQ ID NO: 89, H=C+A, N=A+C+T+G, Y=C+T, K=G+T, R=A+G, W=T+A)), and one primer was designed from the "THC" motif (primer THC: 5'-C GGC ATG GTC GGC TCC HTN ACN CAY TG-3', SEQ ID NO:90, H=C+A, N=A+C+T+G, Y=C+T, K=G+T,

R=A+G, W=T+A); this motif is not universally conserved in PPTases of all organisms).

Using *S. verticillus* chromosomal DNA as template, no amplification product was detected using the THC and the KEA-1 primers. The set of primers THC/KEA-2 successfully amplified a single band of the expected size (about 250 bp), which was gel-purified and cloned. Eight individual clones were sequenced, and all of them resulted to be identical

(except differences due to primer utilization) and highly similar to the putative actinomycete PPTases. The PCR fragment was used as a probe to screen a *S. verticillus* genomic library by colony hybridization. Of the 10,000 colonies screened, 25 positive clones were identified, and then confirmed by Southern analysis to contain the same 4. 6-kb *Bam*HI hybridizing band. The 4. 6-kb DNA fragment was subcloned, and the nucleotide sequence of a 1,761-bp *Bam*HI-*Sal*I region was determined (SEQ ID NO. 3).

Sequence analysis of the *pptA* locus.

The sequence of the 1,761-bp *Bam*HI-*Sal*I fragment was analyzed for coding regions by using the CODONPREFERENCE and TESTCODE programs of the GCG package (Genetics Computer Group, Madison, Wisconsin). Two complete ORFs (*pptA*, *orf3*) and two incomplete ORFs (*orf1*, *orf4*) were identified within the sequenced region (Figure 13). The first ORF from left to right (designated *orf1*) starts out of the analyzed area and ends with a TGA codon at position 248 of the sequenced fragment. Comparison of the deduced product of *orf1* with proteins in databases showed similarities with Rv2795c from *Mycobacterium tuberculosis* (GenBank AL008967) and SC5A7. 22 from *S. coelicolor* (GenBank AL031107), both of unknown function. The second ORF, *pptA*, contains the sequence amplified by PCR and used for the cloning of this locus. It comprises 741 nucleotides, starting with a GTG codon (position 245) which is coupled to the stop codon of *orf1*, and ending with a TAA codon. The starting codon of *pptA* is preceded by a potential ribosomal binding site (RBS), GGGAG. The overall (76. 6%) and third codon position (93. 9%) G+C contents and the codon usage of *pptA* are similar to those found in other *Streptomyces* genes, with the exception of the stop codon (TAA), which is most uncommon in this group of organisms (Wright et al. *Gene* (1992) 113:55-65). The *pptA* gene encodes a protein of 246 amino acids with a predicted molecular mass of 25,619 Da and a pI of 4. 76,

which contains the conserved PPTase motifs. Databases searches with PptA showed significant similarities to the putative actinomycete PPTases (39-52%/48-61% identity/similarity) and to confirmed bacterial PPTases such as EntD from *E. coli* (17%/24% identity/similarity) (Lambalot et al. *Chem. Biol.* (1996) 3:923-936). The third ORF, *orf3*, is separated from *pptA* by an apparently noncoding DNA region of 153 bp, and it is transcribed in opposite and convergent direction with respect to *orf1-pptA*. The gene *orf3* comprises 240 nucleotides, starting with an ATG codon (position 1358) and ending with TGA. The starting codon of *orf3* is preceded by the sequence GAAGG, a potential RBS. The deduced product of *orf3* encodes a protein of 79 amino acids with a predicted mass of 7,555 Da and a pI of 7.17. The Orf3 protein shows similarities to the N-terminal region of SC5H1.35c, a protein of unknown function from *S. coelicolor* (GenBank AL049863). Analysis of Orf3 with the SignalP program (Nielsen et al. *Protein Engineer.* (1997) 10:1-6) predicts an N-terminal signal peptide which would be cleaved between residues 27 and 28 (ALA-DS), suggesting that the mature protein (52 amino acids, 5,099 Da, pI 4.31) would be secreted. Between *orf3* and *orf4* there is an apparently noncoding region of 251 nucleotides. The *orf4* gene is transcribed in opposite and divergent direction with respect to *orf3*. It starts with an ATG codon at position 1610, preceded by a potential RBS (GGAGG), and ends out of the sequenced fragment. The deduced protein product (50 amino acids) of the incomplete *orf4* contains a potential NAD/FAD binding motif, GXGX₂GX₃GX₆G (Scrutton et al. *Nature* (1990) 343:38-43), showing low similarities to diverse oxidoreductases.

Heterologous expression and biochemical characterization of PptA.

In order to test if *pptA* actually encodes a functional PPTase, we decided to overproduce and purify the PptA protein, and assay its catalytic competence on putative substrate proteins or domains. The *pptA* coding sequence was amplified by PCR and cloned into the T5-promoter-based pQE-70 vector, yielding plasmid pQEPT, in such a way that a hexahistidine tag would be added at the C-terminus of the protein. Expression of the pQEPT construct in *E. coli* M15(pREP4) resulted in the overproduction of soluble His-tagged PptA which was readily purified by affinity chromatography on Ni-NTA agarose under non-denaturing conditions (FIGURE). Because *pptA* belongs, by sequence similarity, to the subfamily of PPTases involved in nonribosomal peptide synthesis, we first assayed its activity using two different apo-PCPs as protein substrates. The first one, BlmI, has been previously characterized in our laboratory as a discrete peptidyl carrier protein, or type II PCP, whose gene is found within the bleomycin-biosynthesis gene cluster of *S. verticillius*

(Du et al. *Chem. Biol.* (1999) 6:507-517). For the second PCP substrate we used BlmX, a bimodular NRPS protein encoded in the same cluster (Fig. 2), as a source of a type I PCP, i.e. a PCP included in a multidomain NRPS. For the production of this type I PCP, we amplified by PCR a 1,898 bp fragment encoding the adenylation and PCP domains from the second module of BlmX. This DNA fragment was cloned into pMAL-c2x to yield pMAL1617, in which the type I PCP would be produced as a maltose-binding protein (MBP) fusion, MBImX-2, with a predicted molecular mass of 108.5 kDa. Introduction of pMAL1617 in *E. coli* TB1 resulted in good overproduction of MBImX-2, about 40% soluble, which was purified by affinity chromatography using amylose resin. To test the PPTase activity, we incubated the purified PptA with BlmI and MBImX-2 as putative protein substrates in the presence of (³H)-(pantetheinyl)-CoASH, and the tritiated products were subjected to SDS electrophoresis and autoradiography. The well-characterized PPTase Sfp from *B. subtilis*, which exhibits a broad specificity for its protein substrate (Quadri et al. *Biochemistry* (1998) 37:1585-1595), was included as a positive control. In these experiments PptA exhibited a robust phosphopantetheinylation activity on both BlmI and MBImX-2. Having demonstrated that PptA does in fact have PPTase activity on both type I and type II PCP substrates from nonribosomal peptide synthetases, we then proceeded to test two different acyl-carrier proteins (ACPs) as potential substrates. The first one, BlmVIII, is a monomodular multidomain polyketide synthase (PKS) which is encoded in the bleomycin-biosynthesis gene cluster of *S. verticillus* (Fig. 2). BlmVIII contains an ACP domain at its C-terminus, that is a type I ACP. For the second ACP substrate we used TcmM, a type II acyl carrier protein involved in the biosynthesis of the aromatic polyketide tetracenomycin C in *S. glaucescens* (Shen et al. *J. Bacteriol.* (1992) 174:3818-3821; Bao et al. *Biochemistry* (1998) 37: 8132-8138). For the production of TcmM, its coding sequence was transferred from a construct previously made in pET-22b (Gehring et al. *Chem. Biol.* (1997) 4:17-24) into the pET-28a vector to yield pET28a-TcmM, in such a way that a hexahistidine tag should be added at both the N-terminus and the C-terminus of the protein. Plasmid pET28a-TcmM was introduced into *E. coli* BL21(DE3), and TcmM was easily purified by affinity chromatography using Ni-NTA resin. In vitro phosphopantetheinylation assays were performed as before, but using BlmVIII and TcmM as protein substrates, and PptA was able to posttranslationally modified both ACP substrates.

The *pptA* gene is not clustered to the bleomycin-biosynthesis locus.

Some bacterial PPTase genes have been found clustered, or close, to their respective “partner” NRPS genes: *entD* {enterobactin (Coderre et al. *J. Gen. Microbiol.* (1989) 135:3043-3055)}, *sfp* {surfactin (Cosmina et al. *Mol. Microbiol.* (1993) 8:821-831)}, *gsp* {gramicidin (Borchert et al. *J. Bacteriol.* (1994) 176:2458-2462)}, *bli* {bacitracin (Gaidenko et al. *Biotechnologia* (1992) 13-19)}, *lpa-14* {iturin (Huang et al. *J. Ferment. Bioeng.* (1993) 76:445-450)}. To test the possible clustering of *pptA* to the bleomycin-biosynthesis (*blm*) locus, PCR reactions were performed using the THC/KEA-2 primers on several overlapping cosmid clones spanning the *blm* locus plus 30-40 kb upstream and downstream of its putative limits. No amplification product could be obtained in these reactions, showing that the *pptA* gene is not clustered with the *blm* locus.

Discussion

It has been suggested that in organisms containing multiple phosphopantetheine-requiring pathways, each pathway has its own posttranslational modifying activity (Walsh et al. *Curr. Opin. Chem. Biol.* (1997) 1:309-315). Our group has found that *S. verticillus* ATCC15003 contains several PKS and NRPS gene clusters, one of them being responsible for bleomycin production (a hybrid NRPS/PKS system) (Shen et al. *Bioorg. Chem.* (1999) 27:155-171; Du et al. *Chem. Biol.* (1999) 6:507-517). This suggested that the gene encoding the PPTase for the BLM NRPS could be also clustered, or close, to the NRPS genes. However, we have not found this gene after sequencing almost the whole *blm* NRPS locus. Because having this gene could be important for us in order to express functional NRPS modules from the *blm* cluster, we decided to clone the PPTase gene. Additionally, if the “one NRPS cluster - one PPTase” hypothesis was true, it seemed possible to use PPTase sequences as a new kind of probe to clone novel NRPS clusters.

We know that in *S. verticillus* there are several NRPS locus (maybe four), so we expected several “PCP-type” PPTases. However we have amplified only one, and it does not seem to be closely linked to any of the NRPS loci. Interestingly in the actinomycete *Mycobacterium tuberculosis*, whose genome is fully sequenced, there is only one PCP-type PPTase gene, which is not clustered with any of the two NRPS loci present in this organism (Quadri et al, *Chem. Biol.* (1998) 5:631-645). These and other indirect evidences suggest that the idea of cluster-specific PPTases is not the general rule at all but most probably the exception, especially in organisms containing multiple NRPS clusters. And there are strong evidences that at least some PCP-type PPTases can posttranslationally modify PCPs from

different clusters and even different organisms (Quadri et al, *Chem. Biol.* (1998) 5:631-645; Gehring et al, *Biochemistry* (1998) 37:11637-11650). It is most likely that there is only one PCP-type PPTase in *S. verticillus* and that its gene is not necessarily clustered to any of the NRPS loci.

- 5 Biochemical characterization of the purified PptA protein confirmed not only its PPTase activity but also its broad specificity, comparable to that of Sfp. Different apo-PCPs (type I and type II) and a type-I apo-ACP from the bleomycin synthetase, and the type-II apo-ACP from the tetracenomycin PKS of *Streptomyces glaucescens* were efficiently used as substrates by PptA. These results suggest PptA as a good candidate for heterologous
- 10 coexpression with NRPS and PKS genes to overproduce active holo-synthase enzymes.

- It is understood that the examples and embodiments described herein are for illustrative purposes only and that various modifications or changes in light thereof will be suggested to persons skilled in the art and are to be included within the spirit and purview of
- 15 this application and scope of the appended claims. All publications, patents, and patent applications cited herein are hereby incorporated by reference in their entirety for all purposes.

CLAIMS

What is claimed is:

1. An isolated nucleic acid comprising a nucleic acid selected from the group consisting of
 - 5 a nucleic acid encoding any one of *Blm* open reading frames (ORFs) 8 through 41;
 - a nucleic acid encoding a polypeptide encoded by any one of *Blm* open reading frames (ORFs) 8 through 41; and
 - a nucleic acid amplified by polymerase chain reaction (PCR) using
- 10 any one of the primer pairs identified in Table II and the nucleic acid of a bleomycin-producing organism as a template.
2. The isolated nucleic acid of claim 1, wherein said nucleic acid comprises a nucleic acid encoding at least two open reading frames selected from the group consisting of *Blm* open reading frames 8 through 41.
- 15 3. The isolated nucleic acid of claim 1, wherein said nucleic acid comprises a nucleic acid encoding at least three open reading frames selected from the group consisting of *Blm* open reading frames 8 through 41.
4. The isolated nucleic acid of claim 1, wherein said nucleic acid comprises a nucleic acid encoding a C domain lacking one or more His residues of the conserved HHxxxDG active site for transpeptidation.
- 20 5. The isolated nucleic acid of claim 1, wherein said nucleic acid comprises a nucleic acid encoding a protein encoded by a gene selected from the group consisting of *blmI*, *blmII*, and *blmXI*.
6. An isolated nucleic acid comprising a nucleic acid encoding a module
- 25 comprising two or more catalytic domains of a protein encoded by a nucleic acid of a bleomycin gene cluster wherein said catalytic domains are selected from the group consisting of a condensation (C) domain, an adenylation (A) domain, a peptidyl carrier protein (PCP) domain, a condensation/cyclization domain (Cy), an acyl-carrier protein (ACP)-like domain,

an oxidization domain (Ox), a ketoacyl synthase (KS) domain, an acetyl transferase (AT) domain, a ketoreductase (KR) domain, and a methyltransferase (MT) domain.

7. The isolated nucleic acid of claim 6, wherein said nucleic acid comprises a nucleic acid encoding one or more proteins comprising a module selected from the group consisting of NRPS-0, NRPS-1, NRPS-2, NRPS-3, NRPS-4, NRPS-5, NRPS-6, NRPS-7, NRPS-9, and PKS.

8. The isolated nucleic acid of claim 7, wherein said nucleic acid comprises an open reading frame from SEQ ID NO: 1, SEQ ID NO: 2, or SEQ ID NO: 3.

9. An isolated nucleic acid comprising a nucleic acid encoding a protein encoded by a gene from a BLM gene cluster.

10. The nucleic acid of claim 9, wherein said nucleic acid comprises a nucleic acid encoding a protein encoded by a gene selected from the group consisting of *blmI*, *blmII*, and *blmXI*.

11. The nucleic acid of claim 9, wherein said nucleic acid comprises a nucleic acid encoding a protein encoded by a gene selected from the group consisting of *blmIII*, *blmIV*, *blmV*, *blmVI*, *blmVII*, *blmIX*, and *blmX*.

12. The nucleic acid of claim 9, wherein said nucleic acid comprises a nucleic acid encoding a protein encoded by *blmVIII*.

13. The nucleic acid of claim 9, wherein said nucleic acid comprises a nucleic acid selected from the group consisting of *blmI*, *blmII*, and *blmXI*.

14. The nucleic acid of claim 9, wherein said nucleic acid comprises a nucleic acid selected from the group consisting of *blmIII*, *blmIV*, *blmV*, *blmVI*, *blmVII*, *blmIX*, and *blmX*.

15. The nucleic acid of claim 9, wherein said nucleic acid comprises *blmVIII*.

16. An isolated nucleic acid comprising a nucleic acid that encodes a protein comprising at least one catalytic domain selected from the group consisting of a condensation (C) domain, an adenylation (A) domain, a peptidyl carrier protein (PCP)

domain, a condensation/cyclization domain (Cy), an acyl-carrier protein (ACP)-like domain, an oxidation domain (Ox), a ketoacyl synthase (KS) domain, an acetyl transferase (AT) domain, a ketoreductase (KR) domain, and a methyltransferase (MT) domain, and that hybridizes to a nucleic acid selected from the group consisting of *orf8*, *orf9*, *orf10*, *orf11*,
5 *orf12*, *orf13*, *orf14*, *orf15*, *orf16*, *orf17*, *orf18*, *orf19*, *orf20*, *orf21*, *orf22*, *orf23*, *orf24*, *orf25*, *orf26*, *orf27*, *orf28*, *orf29*, *orf30*, *orf31*, *orf32*, *orf33*, *orf34*, *orf35*, *orf36*, *orf37*, *orf38*, *orf39*, *orf40*, and *orf41* under stringent conditions.

17. The nucleic acid of claim 16, wherein said isolated nucleic acid comprises a nucleic acid encoding a module.

10 18. The nucleic acid of claim 16, wherein said isolated nucleic acid comprises a nucleic acid encoding a BLM gene.

19. An isolated nucleic acid comprising a nucleic acid selected from the group consisting of consisting of *orf8*, *orf9*, *orf10*, *orf11*, *orf12*, *orf13*, *orf14*, *orf15*, *orf16*, *orf17*, *orf18*, *orf19*, *orf20*, *orf21*, *orf22*, *orf23*, *orf24*, *orf25*, *orf26*, *orf27*, *orf28*, *orf29*,
15 *orf30*, *orf31*, *orf32*, *orf33*, *orf34*, *orf35*, *orf36*, *orf37*, *orf38*, *orf39*, *orf40*, and *orf41*, or an allelic variant thereof.

20. The nucleic acid of claim 19, wherein said nucleic acid comprises a nucleic acid that is a single nucleotide polymorphism (SNP) of a nucleic acid selected from the group consisting of consisting of *orf8*, *orf9*, *orf10*, *orf11*, *orf12*, *orf13*, *orf14*, *orf15*,
20 *orf16*, *orf17*, *orf18*, *orf19*, *orf20*, *orf21*, *orf22*, *orf23*, *orf24*, *orf25*, *orf26*, *orf27*, *orf28*, *orf29*, *orf30*, *orf31*, *orf32*, *orf33*, *orf34*, *orf35*, *orf36*, *orf37*, *orf38*, *orf39*, *orf40*, and *orf41*.

21. An isolated gene cluster comprising open reading frames encoding polypeptides sufficient to direct the assembly of a bleomycin.

22. An isolated multi-functional protein complex comprising both a
25 polyketide synthase (PKS) and a peptide synthetase (NRPS).

23. An isolated nucleic acid encoding a multi-functional protein complex comprising both a polyketide synthase (PKS) and a peptide synthetase (NRPS).

24. An isolated polypeptide comprising a catalytic domain encoded by a nucleic acid of a bleomycin gene cluster wherein said nucleic acid comprises a nucleic acid selected from the group consisting of

5 through 41; and

a nucleic acid encoding any one of *Blm* open reading frames (ORFs) 8 through 41; and
a nucleic acid amplified by polymerase chain reaction (PCR) using any one of the primer pairs identified in Table II.

25. The polypeptide of claim 25, wherein said polypeptide comprises an enzymatic domain selected from the group consisting of a condensation (C) domain, an adenylation (A) domain, a peptidyl carrier protein (PCP) domain, a condensation/cyclization domain (Cy), an acyl-carrier protein (ACP)-like domain, an oxidation domain (Ox), a ketoacyl synthase (KS) domain, an acetyl transferase (AT) domain, a ketoreductase (KR) domain, and a methyltransferase (MT) domain.

26. The polypeptide claim 25, wherein the nucleic acid of a bleomycin gene cluster comprises a nucleic acid encoding at least two open reading frames selected from the group consisting of *Blm* open reading frames 8 through 41.

27. The polypeptide claim 25, wherein said nucleic acid of a bleomycin gene cluster comprises a nucleic acid encoding at least three open reading frames selected from the group consisting of *Blm* open reading frames 8 through 41.

28. The polypeptide claim 25, wherein said polypeptide comprises a C domain lacking one or more His residues of the conserved HHxxxDG active site for transpeptidation.

29. The polypeptide claim 25, wherein said polypeptide is a polypeptide encoded by a gene selected from the group consisting of *blmI*, *blmII*, and *blmXI*.

30. An isolated polypeptide comprising a module comprising two or more catalytic domains of a protein encoded by a nucleic acid of a bleomycin gene cluster wherein said catalytic domains are selected from the group consisting of a condensation (C) domain, an adenylation (A) domain, a peptidyl carrier protein (PCP) domain, a condensation/cyclization domain (Cy), an acyl-carrier protein (ACP)-like domain, an

oxidization domain (Ox), a ketoacyl synthase (KS) domain, an acetyl transferase (AT) domain, a ketoreductase (KR) domain, and a methyltransferase (MT) domain.

31. The polypeptide of claim 30, wherein said polypeptide comprises a module selected from the group consisting of NRPS-0, NRPS-1, NRPS-2, NRPS-3, NRPS-4, NRPS-5, NRPS-6, NRPS-7, NRPS-9, and PKS.

32. An isolated polypeptide encoded by a gene from a BLM gene cluster.

33. The polypeptide of claim 32, wherein polypeptide is encoded by a gene selected from the group consisting of *blmI*, *blmII*, and *blmXI*.

34. The polypeptide of claim 32, wherein said nucleic acid comprises a nucleic acid encoding a protein encoded by a gene selected from the group consisting of *blmIII*, *blmIV*, *blmV*, *blmVI*, *blmVII*, *blmIX*, and *blmX*.

35. The polypeptide of claim 32, wherein polypeptide is encoded by *blmVIII*.

36. An isolated polypeptide comprising a module wherein said module is specifically bound by an antibody that specifically binds to a BLM module selected from the group consisting of NRPS-0, NRPS-1, NRPS-2, NRPS-3, NRPS-4, NRPS-5, NRPS-6, NRPS-7, NRPS-9, and PKS.

37. The polypeptide of claim 36, wherein said polypeptide is specifically bound by an antibody that specifically binds to a polypeptide encoded by a gene selected from the group consisting of *blmI*, *blmII*, *blmXI*, *blmIII*, *blmIV*, *blmV*, *blmVI*, *blmVII*, *blmIX*, *blmX*, and *blmVIII*.

38. An isolated polypeptide comprising a polypeptide encoded an open reading frame of a nucleic acid selected from the group consisting of SEQ ID NO:1, SEQ ID NO:2, and SEQ ID NO:3, or an allelic variant thereof.

39. The polypeptide of claim 38, wherein said nucleic acid comprises a single nucleotide polymorphism (SNP) of an open reading of a nucleic acid selected from the group consisting of SEQ ID NO:1, SEQ ID NO:2, and SEQ ID NO:3.

40. An expression vector comprising a nucleic acid of any one of claims 1 through 23.

41. A host cell transformed with an expression vector of claim 40.

42. The host cell of claim 41, wherein said cell is transformed with an exogenous nucleic acid comprising a gene cluster encoding polypeptides sufficient to direct the assembly of a bleomycin or bleomycin analog.

43. The cell of claim 41, wherein said cell is a bacterial cell.

44. The cell of claim 43, wherein said cell is a *Streptomyces* cell.

45. The cell of claim 41, wherein said cell is a eukaryotic cell.

46. A method of chemically modifying a biological molecule, said method comprising contacting a biological molecule that is a substrate for a polypeptide encoded by one or more bleomycin biosynthesis gene cluster open reading frames with the polypeptide encoded by one or more bleomycin biosynthesis gene cluster open reading frames, whereby said polypeptide chemically modifies said biological molecule.

47. The method of claim 46, wherein said method comprising contacting said biological molecule with at least two different polypeptides encoded by *blm* gene cluster open reading frames.

48. The method of claim 46, wherein said method comprising contacting said biological molecule with at least three different polypeptides encoded by *blm* gene cluster open reading frames.

49. The method of claim 46, wherein said contacting is in a host cell.

50. The method of claim 49, wherein said host cell is a bacterium.

51. The method of claim 46, wherein said contacting *ex vivo*.

52. The method of claim 46, wherein said biological molecule is an endogenous metabolite produced by said host cell.

53. The method of claim 46, wherein said biological molecule is an exogenous supplied metabolite.

54. The method of claim 46, wherein said host cell is a eukaryotic cell.

55. The method of claim 54, wherein said eukaryotic cell is selected from the group consisting of a mammalian cell, a yeast cell, a plant cell, a fungal cell, and an insect cell.

56. The method of claim 46, wherein said biological molecule is an amino acid and said polypeptide is a peptide synthetase.

57. The method of claim 46, wherein said polypeptide is a methyl transferase.

58. A method of coupling a first amino acid to a second amino acid, said method comprising contacting the first and second amino acid with a recombinantly expressed bleomycin nonribosomal peptide synthetase (NRPS).

59. The method of claim 64, wherein said NRPS is selected from the group consisting of NRPS-5, NRPS-4, NRPS-3, NRPS-9, NRPS-8, and NRPS-7.

60. The method of claim 64, wherein said NRPS is selected from the group consisting of NRPS-6, NRPS-2, NRPS-1, and NRPS-0.

61. The method of claim 64, wherein said contacting is in a host cell.

62. A method of coupling a first fatty acid to a second fatty acid, said method comprising contacting the first and second fatty acids with a recombinantly expressed bleomycin polyketide synthase (PKS).

63. The method of claim 62, said contacting is in a host cell.

64. A method of producing a bleomycin or bleomycin analog, said method comprising:

providing a cell transformed with an exogenous nucleic acid comprising a bleomycin gene cluster encoding polypeptides sufficient to direct the assembly of said bleomycin or bleomycin analog;

culturing the cell under conditions permitting the biosynthesis of bleomycin or bleomycin analog; and
isolating said bleomycin or bleomycin analog from said cell.

65. An isolated nucleic acid comprising a nucleic acid encoding a phosphopantetheinyl transferase said nucleic acid encoding a phosphopantetheinyl transferase being selected from the group consisting of:
a nucleic acid encoding the protein encoded by the nucleic acid of SEQ ID NO:3;
a nucleic acid amplified by polymerase chain reaction (PCR) using primers that specifically amplify ORF 41 (primers: SEQ ID NO:71 and SEQ ID NO:72) and *Streptomyces* nucleic acid as a template;
a nucleic acid encoding a polypeptide having phosphopantetheinyl transferase activity where said nucleic acid specifically hybridizes to the nucleic acid of SEQ ID NO: 3 under stringent conditions.
66. The nucleic acid of claim 65, said nucleic acid comprising a nucleic acid of SEQ ID NO:3.
67. A polypeptide comprising a phosphopantetheinyl transferase encoded by SEQ ID NO:3.
68. A vector comprising the nucleic acid of claim 66.
69. A cell transfected with the vector of claim 68.
70. A method of converting an apo-carrier protein to a holo-carrier protein comprising reacting said apo-carrier protein with a recombinant phosphopantetheinyl transferase encoded by SEQ ID NO:3 and coenzyme A thereby producing a holo-carrier protein.
71. A cell comprising a modified bleomycin gene cluster nucleic acid, said cell producing elevated amounts of bleomycin as compared to the wild type cell.
72. The cell of claim 71, wherein said cell overexpresses a resistance gene from the bleomycin gene cluster.

73. The cell of claim 72, wherein said resistance gene is a gene listed in Table III.

BLEOMYCIN GENE CLUSTER COMPONENTS AND THEIR USES

ABSTRACT OF THE DISCLOSURE

This invention provides detailed sequence analysis and characterization of the gene cluster responsible for the synthesis of bleomycin in *Streptomyces verticillus*. The bleomycin gene cluster provides the first hybrid polyketide synthase/nonribosomal peptide synthetase pathway and elucidation of the various modules and enzymatic domains characterizing the pathway provides convenient synthetic routes for bleomycins, bleomycin analogs, and various other polyketides.

10

15

20

file: c:_docs\2500 uc ott\125us2\2500.125wo0 blm.ap1.doc

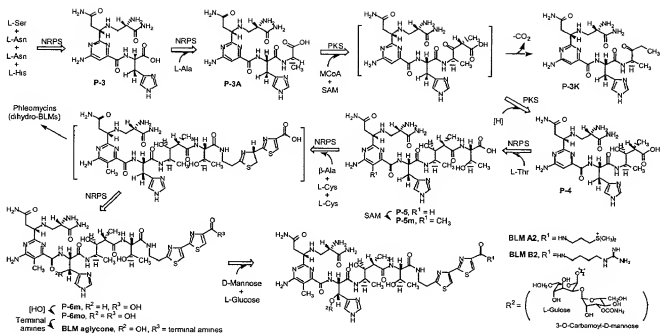


Fig. 1A

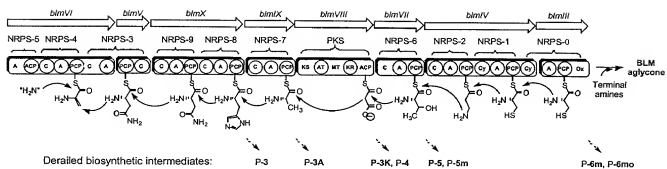


Fig. 1B

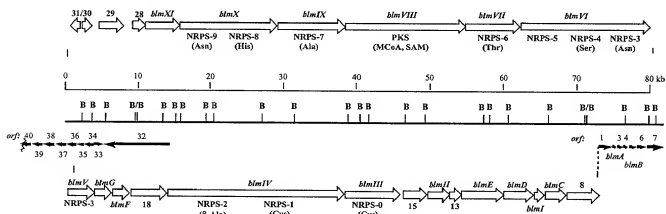


Fig. 2

Fig. 3A

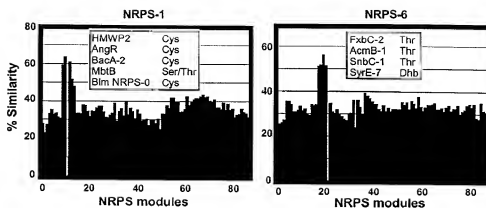


Fig. 3B

NRPS module	Substrate	Residues (PheA numbering) (16)							
		236	239	278	299	301	322	330	331
HMWP2	Cys	L	Y	N	M	S	M	I	W
AngR	Cys	L	Y	N	M	S	M	I	W
BacA-2	Cys	L	Y	N	L	S	L	I	W
MbtB	Ser/Thr	M	L	N	A	G	L	V	H
Blm NRPS-0	Cys	L	Y	H	L	G	L	P	W
Blm NRPS-1	Cys	L	Y	N	L	S	L	I	W
SyrE-7	Dhb	F	W	N	V	G	M	V	H
AcmbB-1	Thr	F	W	N	V	G	M	V	H
SmbC-1	Thr	F	W	N	I	G	M	V	H
FxbC-2	Thr	F	W	N	V	G	M	V	H
Blm NRPS-6	Thr	F	W	S	V	G	M	I	H

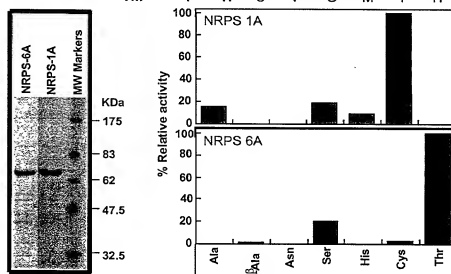


Fig. 3C

Fig. 3D



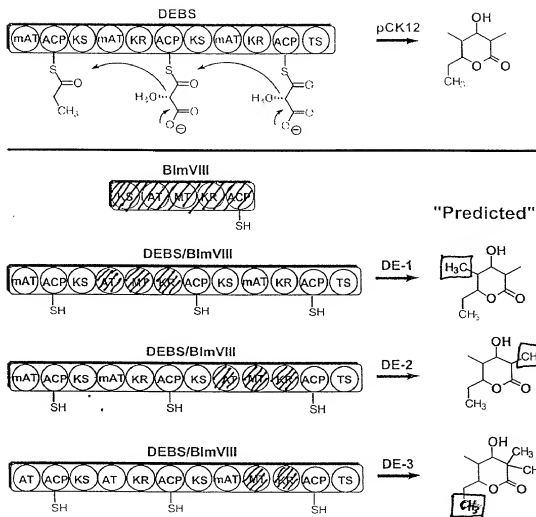


Fig. 5

Fig. 6A

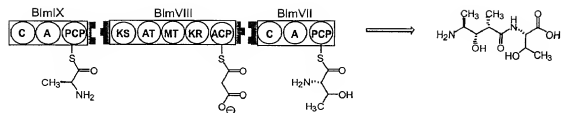


Fig. 6B

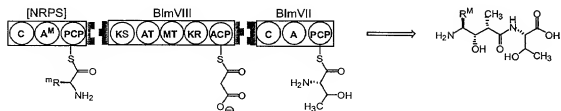


Fig. 6C

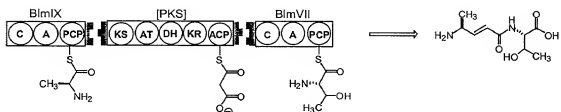


Fig. 6D

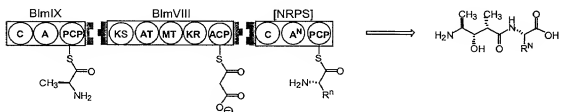


Fig. 6E

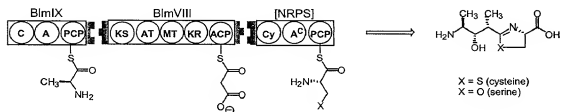
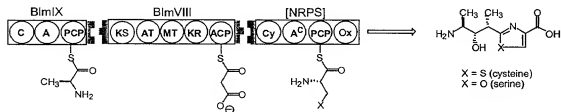


Fig. 6F



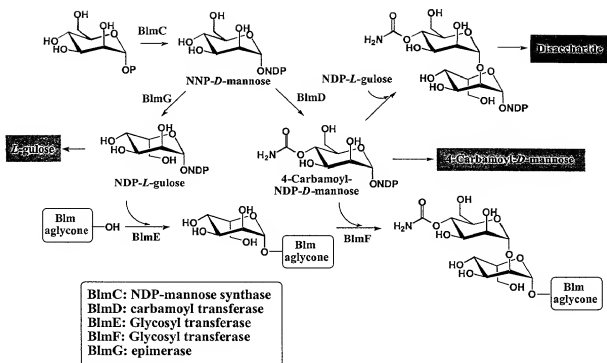


Fig. 7

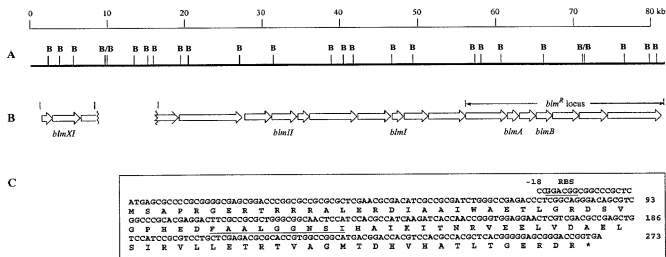


Fig. 8

Downloaded from www.jstor.org

Grs-2 3045-ISIGTEYVAPRTMLEGKLEELWKDVGLQRVTFDDFTI HGHSL-3089
 Srfa-3 960-DQLAEWIGPQNEMETTAQIMSEVGRKQI LDDFAHGHSL-1004
 Vir-S 557-GRSVEGRGVPTPQQEILASLFAEVGLSKV LLEDHGHGHSL-601
 Saf-B 1661-LDPGQDYLAPNNELEARTAAATWEGLERRERVV LSSDDHGHGHSL-1705
 BlmI 1-MSAPRGERTRRALERDLAATWAETSGRDSVP LELAAAGGNLT-45
 consensus 1-1 g eyvapR 1e ia iw evLgr rvGiHddff lGGhsl- 45

Grs-2 3090-KMAVISQVHKECQTEVPLEVLESTPIQGLAKYIEETDTEQYMA-3134
 Srfa-3 1005-KMTAVPH.QQEHGIDLVPVKLLDFAPIAGISAYLKNGGSDGLQD-1048
 Vir-S 602-LTRLTSKIRTVLGAEIAVEDDFAATVEALAEETLEEAREVRPAL-646
 Saf-B 1706-LATRLATLAATIQVQAGVTVSEHRAVAAQAAHFTQATKTHQAH-1750
 BlmI 46-HIKITNEVEELVDAELSIKRVLETRVYAGMTDHFVHATLTGERDR-90
 consensus 46-kAmrv srV 1 ev vivife pTvagla i g t - 90

Fig. 9

100
 101
 102
 103
 104
 105
 106
 107
 108
 109
 110
 111
 112
 113
 114
 115
 116
 117
 118
 119
 120
 121
 122
 123
 124
 125
 126
 127
 128
 129
 130
 131
 132
 133
 134
 135
 136
 137
 138
 139
 140
 141
 142
 143
 144
 145
 146
 147
 148
 149
 150
 151
 152
 153
 154
 155
 156
 157
 158
 159
 160
 161
 162
 163
 164
 165
 166
 167
 168
 169
 170
 171
 172
 173
 174
 175
 176
 177
 178
 179
 180
 181
 182
 183
 184
 185
 186
 187
 188
 189
 190
 191
 192
 193
 194
 195
 196
 197
 198
 199
 200
 201
 202
 203
 204
 205
 206
 207
 208
 209
 210
 211
 212
 213
 214
 215
 216
 217
 218
 219
 220
 221
 222
 223
 224
 225
 226
 227
 228
 229
 230
 231
 232
 233
 234
 235
 236
 237
 238
 239
 240
 241
 242
 243
 244
 245
 246
 247
 248
 249
 250
 251
 252
 253
 254
 255
 256
 257
 258
 259
 260
 261
 262
 263
 264
 265
 266
 267
 268
 269
 270
 271
 272
 273
 274
 275
 276
 277
 278
 279
 280
 281
 282
 283
 284
 285
 286
 287
 288
 289
 290
 291
 292
 293
 294
 295
 296
 297
 298
 299
 300
 301
 302
 303
 304
 305
 306
 307
 308
 309
 310
 311
 312
 313
 314
 315
 316
 317
 318
 319
 320
 321
 322
 323
 324
 325
 326
 327
 328
 329
 330
 331
 332
 333
 334
 335
 336
 337
 338
 339
 340
 341
 342
 343
 344
 345
 346
 347
 348
 349
 350
 351
 352
 353
 354
 355
 356
 357
 358
 359
 360
 361
 362
 363
 364
 365
 366
 367
 368
 369
 370
 371
 372
 373
 374
 375
 376
 377
 378
 379
 380
 381
 382
 383
 384
 385
 386
 387
 388
 389
 390
 391
 392
 393
 394
 395
 396
 397
 398
 399
 400
 401
 402
 403
 404
 405
 406
 407
 408
 409
 410
 411
 412
 413
 414
 415
 416
 417
 418
 419
 420
 421
 422
 423
 424
 425
 426
 427
 428
 429
 430
 431
 432
 433
 434
 435
 436
 437
 438
 439
 440
 441
 442
 443
 444
 445
 446
 447
 448
 449
 450
 451
 452
 453
 454
 455
 456
 457
 458
 459
 460
 461
 462
 463
 464
 465
 466
 467
 468
 469
 470
 471
 472
 473
 474
 475
 476
 477
 478
 479
 480
 481
 482
 483
 484
 485
 486
 487
 488
 489
 490
 491
 492
 493
 494
 495
 496
 497
 498
 499
 500
 501
 502
 503
 504
 505
 506
 507
 508
 509
 510
 511
 512
 513
 514
 515
 516
 517
 518
 519
 520
 521
 522
 523
 524
 525
 526
 527
 528
 529
 530
 531
 532
 533
 534
 535
 536
 537
 538
 539
 540
 541
 542
 543
 544
 545
 546
 547
 548
 549
 550
 551
 552
 553
 554
 555
 556
 557
 558
 559
 560
 561
 562
 563
 564
 565
 566
 567
 568
 569
 570
 571
 572
 573
 574
 575
 576
 577
 578
 579
 580
 581
 582
 583
 584
 585
 586
 587
 588
 589
 590
 591
 592
 593
 594
 595
 596
 597
 598
 599
 600
 601
 602
 603
 604
 605
 606
 607
 608
 609
 610
 611
 612
 613
 614
 615
 616
 617
 618
 619
 620
 621
 622
 623
 624
 625
 626
 627
 628
 629
 630
 631
 632
 633
 634
 635
 636
 637
 638
 639
 640
 641
 642
 643
 644
 645
 646
 647
 648
 649
 650
 651
 652
 653
 654
 655
 656
 657
 658
 659
 660
 661
 662
 663
 664
 665
 666
 667
 668
 669
 670
 671
 672
 673
 674
 675
 676
 677
 678
 679
 680
 681
 682
 683
 684
 685
 686
 687
 688
 689
 690
 691
 692
 693
 694
 695
 696
 697
 698
 699
 700
 701
 702
 703
 704
 705
 706
 707
 708
 709
 710
 711
 712
 713
 714
 715
 716
 717
 718
 719
 720
 721
 722
 723
 724
 725
 726
 727
 728
 729
 730
 731
 732
 733
 734
 735
 736
 737
 738
 739
 740
 741
 742
 743
 744
 745
 746
 747
 748
 749
 750
 751
 752
 753
 754
 755
 756
 757
 758
 759
 760
 761
 762
 763
 764
 765
 766
 767
 768
 769
 770
 771
 772
 773
 774
 775
 776
 777
 778
 779
 780
 781
 782
 783
 784
 785
 786
 787
 788
 789
 790
 791
 792
 793
 794
 795
 796
 797
 798
 799
 800
 801
 802
 803
 804
 805
 806
 807
 808
 809
 810
 811
 812
 813
 814
 815
 816
 817
 818
 819
 820
 821
 822
 823
 824
 825
 826
 827
 828
 829
 830
 831
 832
 833
 834
 835
 836
 837
 838
 839
 840
 841
 842
 843
 844
 845
 846
 847
 848
 849
 850
 851
 852
 853
 854
 855
 856
 857
 858
 859
 860
 861
 862
 863
 864
 865
 866
 867
 868
 869
 870
 871
 872
 873
 874
 875
 876
 877
 878
 879
 880
 881
 882
 883
 884
 885
 886
 887
 888
 889
 890
 891
 892
 893
 894
 895
 896
 897
 898
 899
 900
 901
 902
 903
 904
 905
 906
 907
 908
 909
 910
 911
 912
 913
 914
 915
 916
 917
 918
 919
 920
 921
 922
 923
 924
 925
 926
 927
 928
 929
 930
 931
 932
 933
 934
 935
 936
 937
 938
 939
 940
 941
 942
 943
 944
 945
 946
 947
 948
 949
 950
 951
 952
 953
 954
 955
 956
 957
 958
 959
 960
 961
 962
 963
 964
 965
 966
 967
 968
 969
 970
 971
 972
 973
 974
 975
 976
 977
 978
 979
 980
 981
 982
 983
 984
 985
 986
 987
 988
 989
 990
 991
 992
 993
 994
 995
 996
 997
 998
 999
 1000

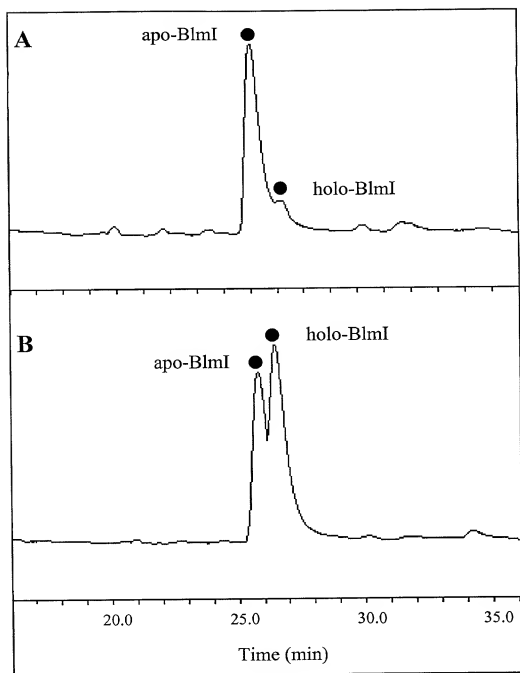


Fig. 10

Nonreiterative Type I Modular Protein Template

Fig. 11A

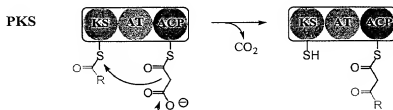
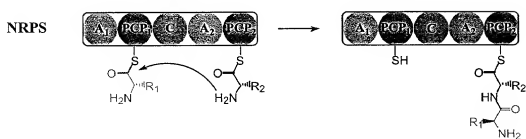


Fig. 11B



Iterative Type II Protein Complex

Fig. 11C

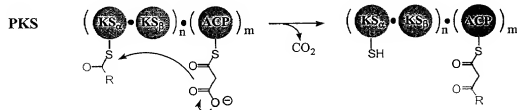
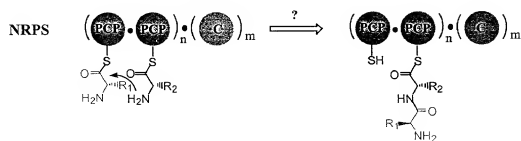


Fig. 11D



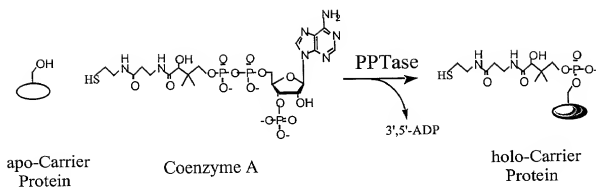


Fig. 12

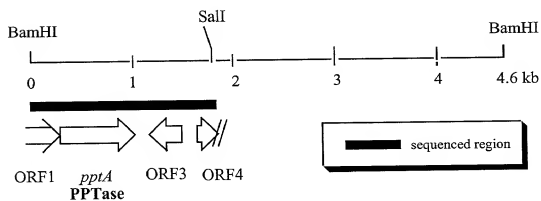


Fig. 13

PATENT APPLICATION DECLARATION

(Attorney's Docket No.: 2500.125US2)

Each of the Applicants named below hereby declares as follows:

1. My residence, post office address and country of citizenship given below are true and correct.

2. I believe I am the original, first and joint inventor of the subject matter which is claimed and for which a patent is sought in the patent application entitled "BLEOMYCIN GENE CLUSTER COMPONENTS AND THEIR USES," Serial No. _____, filed January 5, 2000, and I have reviewed and understand the contents of the specification, including its claims.

3. I acknowledge my duty to disclose to the Office all information known to me to be material to patentability of this application, in accordance with 37 C.F.R. Section 1.56, which is defined on the attached page.

I further declare that all statements made herein of my own knowledge are true and that all statements made on information and belief are believed to be true; and further that these statements were made with the knowledge that willful false statements and the like so made are punishable by fine or imprisonment, or both, under Section 1001 of Title 18 of the United States Code, and that such willful false statements may jeopardize the validity of the application or any patent issuing thereon.

Date: _____

Residence and
Post Office Address:

Ben Shen
1842 Rushmore Lane
Davis, California 95616
(Citizenship: People's Republic of China)

Date: _____

Residence and
Post Office Address:

Liangcheng Du
1301 Orchard Park Q-9
Davis, California 95616
(Citizenship: Peoples Republic of China)

Date: _____

Cesar Sanchez

Residence and
Post Office Address:

(Citizenship: Spain)

Date: _____

Mei Chen

Residence and
Post Office Address:

1301 Orchard Park Q-9
Davis, California 95616
(Citizenship: Peoples Republic of China)

Date: _____

Daniel J. Edwards

Residence and
Post Office Address:

425 Russell Park, Apt. 4
Davis, California 95616
(Citizenship: United States)

Section 1.56 Duty to Disclose Information Material to Patentability.

(a) A patent by its very nature is affected with a public interest. The public interest is best served, and the most effective patent examination occurs when, at the time an application is being examined, the Office is aware of and evaluates the teachings of all information material to patentability. Each individual associated with the filing and prosecution of a patent application has a duty of candor and good faith in dealing with the Office, which includes a duty to disclose to the Office all information known to that individual to be material to patentability as defined in this section. The duty to disclose information exists with respect to each pending claim until the claim is cancelled or withdrawn from consideration, or the application becomes abandoned. Information material to the patentability of a claim that is cancelled or withdrawn from consideration need not be submitted if the information is not material to the patentability of any claim remaining under consideration in the application. There is no duty to submit information which is not material to the patentability of any existing claim. The duty to disclose all information known to be material to patentability is deemed to be satisfied if all information known to be material to patentability of any claim issued in a patent was cited by the Office or submitted to the Office in the manner prescribed by §§ 1.97(b)-(d) and 1.98. However, no patent will be granted on an application in connection with which fraud on the Office was practiced or attempted or the duty of disclosure was violated through bad faith or intentional misconduct. The Office encourages applicants to carefully examine:

- (1) prior art cited in search reports of a foreign patent office in a counterpart application, and
 - (2) the closest information over which individuals associated with the filing or prosecution of a patent application believe any pending claim patentably defines, to make sure that any material information contained therein is disclosed to the Office.
- (b) Under this section, information is material to patentability when it is not cumulative to information already of record or being made of record in the application, and
- (1) It establishes, by itself or in combination with other information, a prima facie case of unpatentability of a claim; or
 - (2) It refutes, or is inconsistent with, a position the applicant takes in:
 - (i) Opposing an argument of unpatentability relied on by the Office, or
 - (ii) Asserting an argument of patentability.

A prima facie case of unpatentability is established when the information compels a conclusion that a claim is unpatentable under the preponderance of evidence, burden-of-proof standard, giving each term in the claim its broadest reasonable construction consistent with the specification, and before any consideration is given to evidence which may be submitted in an attempt to establish a contrary conclusion of patentability.

(c) Individuals associated with the filing or prosecution of a patent application within the meaning of this section are:

- (1) Each inventor named in the application;
 - (2) Each attorney or agent who prepares or prosecutes the application; and
 - (3) Every other person who is substantively involved in the preparation or prosecution of the application and who is associated with the inventor, with the assignee or with anyone to whom there is an obligation to assign the application.
- (d) Individuals other than the attorney, agent or inventor may comply with this section by disclosing information to the attorney, agent, or inventor.

SEQUENCE LISTING

SEQ ID NO: 1 BLM gene cluster ORFS 30 through 8

(note orf 31-40 on sequence 1-18660 are translated on the reverse strand and on a separate file)

18601	ACCCATCTCATAGGTGTACGCGCTGGAGCATTGCGGGCAOGAOGGAAGTTCTCGGTAC	18660
18661	GAGAGCACTGTAAAGCCCGAACCCGCAAGGATGACGAATTCGAAAATTTGTCAGAGTCGCTA	18720
18721	CATGATGGTTCGCGCTGTGCCCGCAGGTAGCCGCGGGCAGCAGCAGAGCTGCCTCGC	18780
18781	GCACCGCGCGGGAGGCCCGGTGAGGCGAGAGGCTGAGGTTCCGTCGCGTTCCGCTGTAT	18840
	M P V P L Y	(orf30)
18841	CAGGCGAAGGCCGAGTTCTTCCGGATGCTGGGGCACCCTGCGCATCCGCTACTGAG	18900
18901	CTGCTGCAGGACGGGCGGATGCCGGTGCCTGATCTGCTGGCGGCGATCGAGATCGAGCCC	18960
18961	TGGGCGCTGCCAGCAGCTGGCGGTTGTCGCGCGCTCGGGCATCGTACCTCCACCCGCG	19020
	S A L S Q Q L A V L R R S G I V T S R S	
19021	ACGGGTTCCACGCTCGTCTACGAGCTGGCCGGTGGCGAGCTGGCGAGCTGATGTCGCGC	19080
	T G S T V V Y E L A G G D V A E L M S A	
19081	GCGCGCCGATCCTGACCGAGATGCTCAATGGGCAGCAGAGCTCTGGAGGAGCTGAGG	19140
	A R R I L T E M L N G Q H E L L E E L R	
19141	GAAGCCGAGGTCAGTCCCGGTGAGCTCCCTCGCCGTCGCGTGGGAGCCCGGTCGCTT	19200
	E A E V S A R *	
	M S S L A V R V G A R V R S	(orf29)
19201	CCGTGCTGCCACCCCGCGCGACCTCGCGGGCATGGCGCGCAGCCCGCAGCTGATCTAC	19260
	V L P T R A D L A G M G R S P R R D L L	
19261	TGGCGGTCGACCGTGGCGATGCTGGCCCTGCGCTCGCCCTCGGATTCGCGCTCTCCT	19320
	A G L T V A I V A L P L A L G F G V S S	
19321	CCGGTCTCGGCGCGGAGCGAGGCTGGCCACCGCGTGGTGGCGCGCGCTGGCCGCGG	19380
	G L G A E A G L A T A V V A G A L A A V	
19381	TATTCGTTGGTGAATCTCCAGGTGTCGCGGCCACCGGCGCATACCGTGGTCTCTG	19440
	F G G S N L Q V S G P T G A M T V V L V	
19441	TGCCCATCGTGCCTCGGTACGGCCCGCGGCTGCTCAGGTGGCCCTGCTGCGCGGAC	19500
	P I V A R Y G P G G V L T V G L L A G L	
19501	TGATGCTGATCGGCTCGCCCTCGCCCGCGCGCGCGCTACATGAGTACGTGCCCGCCC	19560
	M L I A L A L A R A G P R A G P A P	
19561	CGGTGGTGGAGGCTTACCTCGGCATCGCTGCGTATCGGCTGCGAGCAGTGGCGA	19620
	V V E G F T L G I A C V I G L Q Q V P N	
19621	ACGCCCTGGGAGTCGCCAAGCCGAGGGCGACAAAGTCTCTGCTGCTGACCTGGCGCGGG	19680
	A L G V A K P E G D K V L V V T W R A V	
19681	TGAGACCTTCGCGGGGCGCCCAACTGGACCGCTGCCGACTGGCGGCGAGCGTCCGCG	19740
	E T F A G A P N W T A A G L A A A V A A	
19741	CGGTATGCTGACCGCGCGCGGTTGGCGCGGCTGCTTCCCTTCTCCCTCTCGCGGTGA	19800
	V M L T G A R W R P V V P F S L L A V T	
19801	CGGTGCGCACCGTCTGGCCAGCTGTGCCACTGACGCGCGCCCGCATCGGGGACC	19860
	G A T V V A Q L C H L D A G A R P I G D L	
19861	TGCCCGCGGGCTGCCCGCCCGTGGCTTCTGAGACTGGAGCACTGGGCTCGC	19920
	P A G L P A P S L A F L D L D L G A G C S L	
19921	TGCTGGCGCTCGCGTGGCGGCTGGCGGCCCTTGGCGCTTGGAAATGCTGCTGGCGGT	19980
	L A P A V A V A A L A A L E S L L S A S	

19981 CGGTCGCGGACGGCATGACGGTCGGGCGAGAAGCACGACCCGGACAAAGGAGCTGTTCCGGGC 20040
V A D G M T V G Q K H D P D K E L F G Q

20041 AGGGTCTCGCCAACTGGCCGCCCCGCTGTTCCGGCGCGTCCCGGCCACCGGGCGGATAG 20100
G L A N L A A P L F G G V P A T G A I A

20101 CCGGACCCCGCGTCAACGTCCTGACCGGTGCGAGCTCGCGACTCGGCGGCCCTCACGCAAG 20160
R T A V N V R T G A S S R L A A L T H A

20161 CGCGCATCTCGCCGTCATCGTCTTCGCGCGCGCCCACTGGTCTCCGCGCATCCCCCTGG 20220
C A I L A V I V F A A A P L V S R I P L A

20221 CGCGCATCTCGCGCGTGTCTGATCGGACCGCGATCCGCGATGGTCGAAGTGGGCGAGCTGC 20280
A L A G V L I A T A I R M V E V G S L R

20281 GGGCGATGGCCCGGCCACGCGCTCCGACGGCGCTGGTACTGATCTCACGGCGGTGCGCA 20340
A M A R A R T R S D G L V L I L T A V A T

20341 CCGTGCCCTGGACCTCGTCTACGCGCTCATCTCGGCGTGTGGTCCGCGCGCACTCG 20400
V A L D L V Y A V I I G L L V A G A L A

20401 CCCTCGGCGCGTGGCCAAAGCAGGTCCGCGCTGGACCAAGGTCTCCTTGAAGGAGGACCTGA 20460
L R A V A K Q V R L D Q V S L K E D L T

20461 CGGCGACCAACAGCGCCGAGGAACACGCGCTGCTCGCGGACACATCTGGCGGTACCCGCA 20520
G D H S A E E H A L L A E H I V A Y R I

20521 TCGACGTCCTGCTGTTCTTCGCGCGCGCCACCGTCTCTGCTGGAATCTCGGACGTGCG 20580
D G P L F F A A A H R F L L E L S D V A

20581 CGGACGTGCGGTGGTGTATCTGCGCATGTCCGCGTGACCAACATGGAGCCACCGGCG 20640
D V R V V I L R M S R V T T M D A T G A

20641 CCTCTGCTCTGAAGGACGCGGTACCAAGCTGAACCGCGCGGCATCACCTCTCTGGCTTC 20700
L R L K D A V T K L N R R G I T V L A S

20701 CCGGGGTAGCCCGCGCCAGCGCGCGGTCTCTGACTCCGTGCGCGCCCTCGGTCTGCTCC 20760
G V R P G Q R R V L D S V G A L G L L R

20761 GGGCGCCACCGCGACGACTACACCGGCACTCCGGAAGCATGCGCGCCCGCCGAGCC 20820
A A T G D D Y T G T P E A I A A A R S H

20821 ACCTGACGGCGCGGTGTCTTGGCCCCGCGCTGCCCGGCGCGCTCTCCGTTACCCC 20880
L H G A G V L A P A C P G P P P P P

20881 CACCGTCGCTCCGAGTCCCGACGATGAGGAGCCGACCGAGTCTCTCTCCGTACCCG 20940
P C A P S A R R *

20941 GACACCCACGGTTGCGCGCCCCATGCGCGCGGTCTCTCTGACGCGCCGTCCGCGGCTT 21000

21001 GAGGCGCGGTGGAACGGCTGCGCGCGCGCTCGGCTGATCGCGGTGATCACGCGCC 21060

21061 ATGCGCGGTGGCGCGCGCATCGTGGCGGGACCGTGTTCGCGGCCACCGCGCGG 21120

21121 CGGCGCTGCGCTGGCGGTGGCTGCGCGGTGCTGTAGCGCGCGGGTCCGCGGCGG 21180

21181 GGCCTGTGCTTCTTCCGCGCGTCCGCGGGTGGCGCGCGCGCGCGGTGACAGGGAAT 21240

21241 ATGACCGGAATGGGATGCTCGCGTCCACTCGGCTGTGTTAAGTGCCACGGGGGCTTCC 21300

21301 GACGGCGGTGCGCGCGCGGTTCGCGCGATGATGGTCTGCGCGGTGTGACGCGGG 21360

21361 GAGCCTATGGCACAGGACTGAACGACTGGATCGAGGACGAGTCTGCTCCCTACGAGGAG 21420
M A Q D L N D W I E D E V V P Y E E (orf28)

21421 AAGCCTCTGGAATGGATCTCCAGTACCACTTCTTCCGGAACCGCGCGAGCCGCTAT 21480
K P L E W I S Q Y H F F R D P A R A A Y

21481 GTCGATCACACTACTTCTTCTCACCGGCGATGGCGGATCTACCAAGAAAGTAGT 21540
V D H T Y F F S P A D G A I V Y Q K G V

21541 GATCCCCAGGAGTCGATCATCGACATCAAGGGGAAGCGTACTCGTGGCGCGCGCTCT 21600
D P Q E S I I D I K G K P Y S L A A A L

21601 CGTGACGAATCGTTTCGGTCACCGGTGCTGGTGATCGGCATCTTCATGACCTTCTTCGCAC 21660
R D E S F G H R C L V I G I F M T F F D

21661 GTGCACATCAACCGGATGCTTACCGGGGGCGTCTCTCCTTCGCGCTCAAGAGGCCCATC 21720
V H I N R M P Y G G R L S F A L K E E P

21721 GGGACGTTCAACCTCCCCATGCTGGCCATGGAGCAGGACCTGCTCGAAGCGGCTCCGGGTC 21780
G T F N L P M L A M E Q D L L E R L R V

21781 AATCCGGCTCAACGAGGTATCTGCACCTGAACGAGCGGATGGTCAACCGGCTCGACCGG 21840
N P A H A R Y L H L N E R M V N R V D A

21841 CCGGGCTCCGGGGCCCGTACTGGATGCTCCAGATCGCGCATCAAGACGCTCCCATC 21900
R Y G S Q V D L V I P M A A D R E Y V P

21901 ACCCGGTTCTGCAGACGGAGGAATGTTCCGCTCCAGGGGCGCGCTTCTCCAGATC 21960
T P F C R R Q G M F R S Q G R R F S Q I

21961 CGCTACGGATCGCAGGTGACCTGGTGATCCCGATGGCGGCGAGCCGCGAGTACGTCGCC 22020
R Y G S Q V D L V I P M A A D R E Y V P

22021 GTGAGGCGCGTCCGCGGCACTGTAAGCGGGGCTCGACCCGCTCGTCAAGATCCGGTGG 22080
V E A V G R H V K A G L D P L V K I R W

22081 CGTTGAAGACGCGCTACGAAGCGATGGCGAATGGAGGGACACAGCGTGGGTTTCGTCG 22140
R * M G F R R (orf27)

22141 AGCGCAGAGGGCGCGTGGGCGGGAGCGGGCGCGCGGAGAGCGCCGGTTTCAGGCCGA 22200
A Q R A G G P G A G R R R E S A R F R P D

22201 CCGGCGCTCGGCGCGCGGGGACCGTCCGTTACCCCTGTCGCGCGGGCAGTTGTTTCGAGTG 22260
G P S A P R D R P L P L S A G Q L F E W

22261 GGTGTTTGCAAGCTCGTGCAGCGAGATCTGAGCCACCGCAAGATGTTGCGGGCTCCG 22320
V F D K L V D G D L S H Q C T P I V R L R

22321 CGGCCCGCTGAACACCGCCCGCTCGGATGGCTTACCGCGCGCTGGTGGCGGCCACGA 22380
G P L N T A A L R M A Y A R L V R R H E

22381 GTGCTCGCACCCCGCTTCCCGTGATCGACGGGAGCCCGTGCAGGTGATCGAGGGCAT 22440
C L R T R F P V I D G E P V Q V I E G I

22441 CCGGAAGACGCGGGGGGCGCGTCCGCTCATCGATCTGCGCACCTCCGAGGAGCGCT 22500
G K A A G G P L P L I D L R H L P E A L

22501 TCGCGCGCGAGATCGCGAGGATCCGCGAGGAGCGCTGTCCACGCGGTCCTCCCTTCGA 22560
R A R E I A R I R E E T L S T P V P F D

22561 CAAGCGCGCGCGCTCGCGTGGCGCTGATCCGGCGCGCGCGGAGGACCTCTCTCTCT 22620
K R P P V R V A L I R A A F E H L F L

22621 CGTCGCGATCCCGACATCACC CGGACCTGTGGTCCGCGACCTGCTCAACGAGGCT 22680
V G I P H I T A D L W S A T L L N D E L

22681 CATGGCGCACTACAGGGCGGGGCGAGGGGACTCCCTCCGCGGCCCCACCCCGTGGC 22740
M A H Y R A G A E G T P S R A A P T P V A

22741 CGAGTACCGGACTTCGCGAGTGGCAGCGCGCTGGTGGAAACGGGACCGCAGGAGCG 22800
Q Y A D F A Q W Q R A W N R D R T E R

22801 GGAGGCGCGACGTTGGCGGCGCGCTGGACGGGCTGTCCGCGTGGAACTGCCCTTGA 22860
E A G R W R A R L D G L S A V E L P L D

22861 CCGGCGCGCGCGCGCGCGCGCGGAGTCTCTCTGATCGGGGACACCTTCGAGCGC 22920
R P R P A G R R R D C F L I G D T F D A

22921 CGAAGTGGGAGCGCGCTGGCGCTTGGCAGCACCGCGGACGCTGACGTTGCTGGT 22980
E L S D R L R A L A R T A D V T L Y V V

22981 GCTGCTGGCGCGTTCACCTGGCTGGTGGGCGGATGTCCGGGCGCGGCTGGTGAC 23040
L L A A F H W L V G R M S G A G R L V T

23041 CACCTCGCTGTGGCGCGCGCGGACGGCAGCGCGTACAGGGGATGACGCGCGCGCTTC 23100

T S L V A A R H G S A V Q G M T G P F S
 23101 GGACTACCTGGCCCTGGTGGGGACCTGTGGGGCGATCCGGAACCTCTGGAGTCCCTCGG 23160
 D Y L A L V G D D L S G D P D F L E S L R
 23161 CCGCCTACCGACGAGTGCCTGACCGCCACGACCACCGCGCTCCGTTCTCACAGGT 23220
 R V R D E C L T A H D H Q R L P F S Q V
 23221 CTCTGAAGTCATGGACCCGGACGCGAGTTGACCCCATCCGCTGGAGCAGCTCGGGTT 23280
 L E V M D P G R E L H P H P L E Q L G F
 23281 CAACCTCCACAACATCCCTCCCGCGGTCTGGAGTCTCTCCGGCGAGCTGCTGTCGCG 23340
 N L H N I P P A V M D F S G D V V V S A
 23341 GGTGAACCCGGAGGGGACGACGCGGGAGAGCGGCGACGCGGAGTACGTGCCCTGGACGC 23400
 V N P E G D D G E S G D G E Y V P W T A
 23401 CGACCTGACCTTCGACGTCTACGACTACGGCACCGGCCATATCCGCTTCGACGTGATACT 23460
 D L T F D V Y D Y G T G H M P F D V I L
 23461 CGACCGCGCGTGGCCGATCCGGCGACGCGCCCGGAGTGGCCCGGCACTACCGCTCGGT 23520
 D R R L A D P A T A R E W A G H Y R S V
 23521 GCTCCGTGCGGTCTGTCGCGACCCCGCGGTGCGCTTCTCCGCCCTCGGCACTCTGCTGC 23580
 L R A V V A D P G V R L S A L G T L L S
 23581 CTGCGCGACCGCGTCCGCCACGTCCTTCGGCGCGCGGAGATCGACGTCGCGCGCGT 23640
 L P R P P S A T S F G G R E I D V R R V
 23641 CGAACGCGAGTTGGCGGGGCGCGACGGGATCACCGCGCCTGCGTGGCGTGGCGCCCG 23700
 E R E L A G R D G I T A A L V A V A P R
 23701 GCGCCTGGCCACCGCGGTGCGGTACGGGAACGTGCTGCTACTGCGCGTGGAGGCGAC 23760
 R L A T G L R V R E L V A V C A V E G C
 23761 GCGCGTTCGAACGCGGCCACGACATCCGCGCGCGCTCGGGAGCGCTGCCCGACGG 23820
 F R P N A A H D I R G R L R E R L P D G
 23821 CTGGCGCGACCGCTGTTGTCGACGCGCGCGAGGAGATCCGGAAGCCCTGCGCGC 23880
 W V P T V F V E R P P E E I R K A L A A
 23881 CCGGCGCGCGCGCGGACGCGCGGAGCGCTGCGCGCGCGCGAGGACTGCGTCCCGCT 23940
 R A A G G E R A E P L P P E D C V P L
 23941 TCCGAGGAGGGCGCGCCCCCTCGGACCGTCCGAGCGCGCGCTGCGCGCTCTGGGC 24000
 P E E G R P P S D P S E R R L A A L W A
 24001 CGAGATCCTGGGCGCCCCCGGAAGAGCGTGACCGAGCCCTTCTTCGCGTGGCGCTCAC 24060
 E I L G A P P K S V T E P F F R V G V T
 24061 CGATAAGGACGCCCTCCGCTTCTTGGCGCGGTGGCGGAGGACTTCCGCGTCCCGTGCC 24120
 D K D A L R F L A R V A E D F G V T V P
 24121 CTTCGCGCACTTCTCAGCGCTCCCAACCTGCGTATGGTGAAGGACAAATTGCTGTGAGAA 24180
 F A D F L S A P N L R M V K D N L A E K
 24181 ACGGAGGGTGTAAACGCGCAATGAGTGAAGTGGTAGGGTCGGAATCGAACCGCACTGATCG 24240
 R R V *
 24241 CAATCTTTTCGGTCAGCTGTTCCGATATTCCGCGGCGGTGCGCGCTCTCTCGACCAAG 24300
 24301 GGGCTACGCGGATAAGCGTGCGCCGCCACGGCTGCGTCTCGACGCTTCATCGCGCGG 24360
 24361 TCGGACACTTCTCGGTGCGCTCGGACGCTCAGAGATCAGTGAATGCGCTCGGTGTC 24420
 M P R C A (orf26)
 24421 CGAGGTGCGCTCAGTACTGTGTTCACACAAACGCGCAAGGAGTGAAGCGTATGGAG 24480
 R G A L S T A V H T T R Q G S W N V M E
 24481 ACGGCGAATTCCGGCTATCGGGTCTCACCTCAGCAGCGGCATTTATGGCGCTGCTGACC 24540
 T A N S G Y R V S P Q Q R H L W A M L T
 24541 CGCGGCGGAGCGCGGGCGACGTGCGTTCAACCGCATCGCGCGTGGTGGTGCACCGGTTCC 24600
 R G R D G G R R A F T Q S A V V V D R S

24601 CTGGACGCGCACGCTCTGCGGCGCGCTGGCCTCGTGGTGGCCGCGCCAGCAGCGCGTG 24660
L D A A R L R A A L A S V V A A H E P L

24661 CGGACGACCTTCACCGGTCTCGCGGACGGACCGCGCGGTCCAGGTCTCATGACCCG 24720
R T T F T G L A G R T A P V Q V V H D P

24721 GACGAGCAGCGCTGTCTGCTGCGACTGCGCGCTCTGTCGCGCAGCGCTCGGCGCCG 24780
D E Q P L S V V D L P P S C A D G S G P

24781 GAACCTGGACGAGCTCCGCGTCCGCGAACCGCGCCCTCGACCGCGCGCGCGCGCGCTC 24840
E L D E L R L R E R A A L D P R G G P V

24841 TTCCGGGCGCCCTGGCGCGCGCGCGGAGGACCGGGCGGTGCTGTGTCTCACCGGCAC 24900
F R A A L A R A G E D R A V L V L T A C

24901 GCCCTGTGCGGACCGCTCTCCCTCCGGCTGCTGGCGGGCAGATCTCGCGCGGTAC 24960
A L V A D R L S L R L L A G Q I L A A Y

24961 AGCGGGGAGACCGGTCTCCCGGATGGCGCGCGCTTGCAGTACGCGCAGATTCGCGCGC 25020
S G E T V S P D G P P P L Q Y A D L T A C

25021 TGGCACCAGACCTGCTCAACCGCGAGACGCGCGCGCGCGCGCGCTGGCGCGCC 25080
W H H D L L T A E D A A P D R A H W A A

25081 CACACCGCCACCGCGCGCGCGCGCTCCCGCGCTGCTACGCGCGCGCGCGCGCGCG 25140
H T A T A G T G P L P G V V R P G A A P

25141 GGTCCGTGGCGGGCGCGGAGTGGGAACCTGGCGCGCAACTGGTGGCGGGGATCGACGG 25200
G P W R A R E W E L P A E L V A G I D G

25201 GTCCCGGGAAGCTGTCCACCGATCCCGCCACCGTGTGACGCGCGCTTCGTATCGG 25260
V A G K L S T D P A T V L H A A F R I A

25261 GTCTGGCGGCTCGCGCGGAGGGAACCTGGCGCGCGCGCTCACTCGTACGCGCGTTC 25320
V W R L A G E R N L P V A L T R D G R S

25321 CACCCGAACTCGCACCGCGATCGCGCGCTTCGAGCGTGAGCTCCCGCTGTCCACGAG 25380
H P E L R T A I G A F E R E L P L V H E

25381 ATCCGTCACGAGACCGCGTTCGCGGAATACGCGCGCTCTGACGCGCTCTGCGCGGAG 25440
I R H E T A F A E Y A R A L D A L V A E

25441 GGCAGGAACCTCTCGACCAATGCGACCGGAACTGCTCGGCGAGCTCGACGGCACCGG 25500
G E E L L D H C D P E L L G S L D G T A

25501 GAAGGCGCTGCTTACCTTACCCACACCCAGGCGGAAACACCGGTTCGGCGGGCGCGG 25560
E G P C F T F T H H Q A E T P V R R A G

25561 ATCACTTTACCAACGCTCATCAGGATTCGGGTACGCGGATTCCTCGCGCTGACCGCC 25620
I T F T T V H Q D S G T P I P V R L T A

25621 CGACGCGAGCGCGCGCTCGCGCATGGAACTGGGATACGAGGAGGCGGTATCGACGAG 25680
R R D G A R L R M E L G Y D E G R I D E

25681 ACGTTTCCGAGAACGCGCGCGCTGCGCTCACCCGCAITCTCGAAGGCGTGCTCTCGCC 25740
T F P E N A A A C L T R I L E G V V S A

25741 CCGGAGGCGCGGTTCGGGACATCGCATGTCTCGGACGAGACCGCAGCGGTGCTCGG 25800
P E G P V G D I R M L S D E T A R L R A G

25801 GAAGCGGGGCTGGGCCCCGCGTGGAACTCCCGCAAGCGGTCCAGAACTCTTCGCG 25860
E A G L G P R V E L P G K A V H E L F A

25861 GAGCAGCGCGCGCACACCCCGGGCGGTGCGGTACGCGCGGCGAGGACGCGCTCACG 25920
E Q A A R T P G A V A V S A G E D A L T

25921 TACGCGAACTCGACGAGCGGTCAAACCGCTGGCACACCACTGACGCGGCTCGGGGT 25980
Y A E L D E R S N R L A H H L T G L G V

25981 ACAACCGCGCGCACGTCTGGTCTCGGTGCGCGCTCGCGGAGCTGTCTGTCGGGCTG 26040
T P G R H V V V S V G R S A E L L V G L

26041 CTCGCGGTGCTCAAGCGGCTGGCGCTTCGTCGCCGTCGAGTGGGCTTCCCGCGCAA 26100

L G V L K A G G A F V P V D V G F P R K
 26101 CGGCTGGAGTTCGTGCTCCGGGAGACCGCCCGCGGTCTCTGCTCGCACGCCGGAOSTA 26160
 R L E F V L R E T A A P V L L C T L E F T L L
 26161 CGGGACCGCATCGGCACTCGGACCTCGACACGCGCGGGTGACACCCGTCTCGGCTGGAC 26220
 R D R I G T R T L D D A G V T P V A L D
 26221 GCGGACCGCGCGCATCGCGCACACCCCGCGCGCCACCGGCATCGCCACCCACCC 26280
 A D R R R I A A H P A G P T G I A T T P
 26281 GACGCCCCCGTACGTGCTCTACACCTCGGCGACCGGGAAGCCCAACGGCGGTACGC 26340
 D A P A Y V V Y T S G T T G K P N G V R
 26341 GTCCCGCACCGGGGCTCACCACTACCTACCTGGTGACCGGCGCTTACGACTGAC 26400
 V P H R G L T N Y L T W C T G A Y G L D
 26401 GGGGGCACCGGCACCTGTGCACACCTCCATCAGCTTCGACCTCACCTCCACCCCTG 26460
 G G T G T L V H T S I S F D L T L T T L
 26461 TTCCGCCCCCTGCTCGCGGCGGGCAGGTGCTCATGCTCTCCGAGACCCCGCGGTGACC 26520
 F G P L L A G G Q V V M L S E T A G V T
 26521 GCGCTGATCGCGCGCTCGCTCCCGCGCGACCTCACCTGGTCAAGCTGACCCCGACC 26580
 G L I A A L R S R R D L T L V K L T P T
 26581 CACCTCGACGTCTCAACAGCTGCTCACCCCGACGAGCTCGCGGCGCGGTCTCGCACC 26640
 H L D V V N Q L L T P D E L R G A V R
 26641 CTCGTGTCGCGCGGGAGGCGGTGCGGGCGGAGACCTGGAGCGGTTCGGGCGCTCGGG 26700
 L V V G G E A V R A E S L E P F R A S G
 26701 ACGCGGTGTCACAGGTAACGGGCCAGCGAGACGGTCTGTCGCGAGCGTTCGCGACGCT 26760
 T R V V N E Y G P S E T V V G S V A H
 26761 GTGAGCGCGCCACGCCCCGTACCGGCGCGGTGCCCCGCGCGCGCATCGCCACAC 26820
 V D A A T P R T G P V P I G R P I A N T
 26821 ACCGTCCACGTCTCGACACGCGGCGCGCGGTCCCGCGCGCGGTCTGCGGAGAGCTG 26880
 T V H L L D Q R R R P V P D G V V G E L
 26881 TGGATCGGCGCGCGGTGTCGCGGACGCTACCTGGGGCGCGCGGAACCTCACCGCGAG 26940
 W I G G A G V A D G Y L G R P E L T G E
 26941 CGCTTCCTCCCGACGACTACCGCGGACGCGGGCGGGTCTACCGCACCGGCGACCTG 27000
 R F L P S D Y P P D G G R V Y R T G D L
 27001 GCGCGCGCGCGCGCGGACCGGACCTGGAGTACCTCGGGCGCACGACGCGGAGTGAAG 27060
 A R R R A D G T L E Y L G R T D A Q V K
 27061 ATCCGCGCGCGCGGTGAGCCCGCGGACCGAGGCGGTCTCTGCGCTCCACCCCGGCG 27120
 I R G V R V E P A E T E A V L A S H P G
 27121 GTCGGCGAGGCGTCTGTGGTCCCGCGCTGGACGAGGACCCCGGCGGTCTGTCGCGGCTC 27180
 V G Q A V V V A R L D E D P G R S S P L
 27181 GCGCGGAGCTGACGCTGACCGGCTACCTGGTCCCGCGCGCGGTGCGGCGCGCGG 27240
 A G E L T L T G Y V V P A R G A Q A V
 27241 CACGAGGAGCTCATCGCTACTGCGGGAGCGGCTGCCCGAGCACTTCGTCCGCGCGCTC 27300
 H E E L I A Y C R E R L P E H F V P A V
 27301 CTCGTACCCCTCGACGCGCTGCGCGTACCGGCGCACGCAAGATCGACGCGGTGCGCTG 27360
 L V T L D A L P V T G H G K I D R G A L
 27361 CCCAAGCGCACGCGCGGCGCGGACGCGCGGCTACGTGCGCGCGGCGCACCGTACC 27420
 P K P H A R A R D G A A Y V A P R T A T
 27421 GAGGAGATCTCGCGGCCCGTGCAGAGGTGCTGGGCGTGCAGGCGGTGCGCATCGAC 27480
 E E I L A A T V A K V L G V E R V G I D
 27481 GACAACATCTCGTCTGGGCGGCGACTCCATCCGACGCTCATGGTTCGCGCGCGGCG 27540
 D N Y F V L G G D S I R S V M V A S R A

27541 CAGGCCCGCGGGTCGAGGTCACCGTGGCCGACCTGCACCGGCACCCACCGTCGCGGCC 27600
Q A R G V E V T V A D L H R H P T V R A

27601 TGCGCCCGCACCTTGAAGCCCGCGAGGACCTGCGCGGACGCGCGTCACCGAACCTTC 27660
C A A H L D A R E D L P R T P V T E P F

27661 GCGCTGATCTCCGCGAGGACCGGGCGCTGGTGCAGGACGACCTGAGGACGCGCTCCCG 27720
A L I S A E D R A L V P D D V E D A F F

27721 CTGAACCTGTCTCAGGAAGGATGATCTTCCACCGGACCTTCCGCGGAGGATCGGCGGTC 27780
L N L L Q E G M I F H R D F A A K S A V

27781 TACCAGCCATCGCGTCCGTCGGGCTGCGCGCCCGTTGACCTCGCGCTGCTGCGGATG 27840
Y H A I A S V R L R A P F D D L A V L R M

27841 GTGCTGCGCCAGCTCGTCGAGCGGCACCCGATGCTGCGCACCTCCTTCGACATGAGCCG 27900
V V R Q L V E R H P M L R T S F D M S C R

27901 TTCACGCGCCGCTGCAACTGGTGCACCGGATTCGCGGATCCGCTGCACTACGAGGAC 27960
F S R P L Q L V H R E F A D P L H Y E D

27961 CTGCGCGCAGGAGCGCGAGGAGCAGGACGCGCGCTCGAGGATGGATCGAGCGGAG 28020
L R G R S A E E Q D A R V E E W I E R E

28021 AAGGAACGCGGCTTCGAGCTGCACGAGTTCCGCTGATTCGCTTCATGGCGCAGCGCCCTG 28080
K E R G P E L H E F P L I R F M A Q Q L C

28081 GAGGACGAGCTCTTCCAGTTACCTACGGCTTCCACCAAGAGATCGTGAACGCGCTGGAGC 28140
E D D V F Q F T Y G F H H E I V D G W S

28141 GAAGCCCTGATGATCAGCGAGCTGTTACGCCACTACTTCTCGGTGATCTACAGCAGAGCG 28200
E A L M I T E L F S H Y F S Y I Y D E P

28201 ATGCGGATCAAGCCACCCACGCGCGATGCGCGACGCGCTCGCCCTGAGGCTGGAGGCC 28260
I A I K P P T A G M R D A V A L E L E A

28261 CTGCGGACCCGCGCACTACGAGTTCTGGGACTCTTACCTCGCGACGCCACCTGATG 28320
L A D A C C R N Y E F W D S Y L A D D A T L M

28321 CGGCTGCGCCAGCGCGGACCGGACCCCGGGCGGACGAGGCGGACCGGATGACCGCGC 28380
R L P R P G T G P R A D K G D R D I T R

28381 ATCGCGCTCCCGTCCCAACCGAACTCTCGACGGCTCAAGCGGGTGC CGCGCACCCAC 28440
I A V P V P T E L S D G L K R V A A T H

28441 GCGCTCCCGTGAAGACCGTGCTCCTGGCGCGCACATGTTGGTGATGTCCTCTACGG 28500
A V P L K T V L L A A H M V V M S L Y G G

28501 GGCACGAGGACACCTCACTACACCGTCAACCAAGCGCGCCCGGAGACCGCGGACGGC 28560
G H E D T L T Y T V T N G R P E T A D G

28561 AGCACCGCGATCGGGCTGTTGTCACAGCCTCGCGCTCCGCGTCGCGATGACCGCGGC 28620
S T A I G L F V N S L A L R V R M T T G G

28621 ACCTGGGCGACCTGATCACCGCCACGCTGGAGTCCGAGCGCGCTCGATCGGTACCGT 28680
T W A D L I T A T L E S E R A S M P Y R

28681 CGGCTGCCGATGGCCGAATCAAGCGGCCACGAGGCAACGAAACCCCTGCGGAGACGCTG 28740
R L P M A E L K R H Q G N E P L A E T L

28741 TTCTTCTTCAACCACTACCACTCTTCCAGTGTCTGACCGCTGGATCGAACCGCGGCTC 28800
F F T N Y H V F H V L D R W I D R G V

28801 GGCACGCTGCGCAACGAGCTCTACGGGAGTCCACCTTCCCTTCTCGGATCTTCCGC 28860
G H V A N E L Y G E S T F P F F C G I F R

28861 CTGAACCGGGAGACCGGAGCTGGAGTCCGATCGATACGACAGCTGACGTTCTCC 28920
L N R E T G E L E V R I E Y D S L Q F S

28921 GACGCCCTCATGGAGAGCTCGCGGACAGCTACGCCCGCTCCTCGCGGCGCTGCTGCC 28980
D A L M E S V R D S Y A R V L A A L V A

28981 GACCCGCGGCGCTACGACCGGCAAGGATTCGCTCCGACCGGACCGGCGCGCACTG 29040
D P D G R Y D R H E F R S D R D R A A L

29041 GCCGCTCTCA¹CCCGGGG²CCGAGGCGCGGCGCGCA³CCGGTGCCTGCA⁴CAGAC⁵TGGT⁶G 29100
A V L T R G P E A P A A D R C L H D L V

29101 GCGGACCGGGGCGCGGACCGCCCGGACGCC¹CCGGCGGTCCAGCTGGACAC²CGCTGCTC³A 29160
A D R A A D R P D A P A V L D L D T A G V

29161 AGCTACGGGAGCTCGACCGCGCGCCAA¹CCGGCTGGCC²CACCACCTGCGGT³CGCTCGG⁴C 29220
S Y G E L D R R A N R L A H H L R S L G

29221 ATCGGCGCGGAGAGCGTCGTGCGGCTCCTGCGCGAA¹CGCTCCCTCGCCAGATCATCGGC² 29280
I G P E S V V G V L A E R S L A Q I I G

29281 CTCCTCGCGGTCTCAAGGCGGGGCGCGCTACGTCCCGCTCGACCGCGCCACGCCGAC¹ 29340
L L A V L K A G A A Y V P L D P A L D G S

29341 GAGCGCCTCGCGCGCTCATGCGGGAGCGGGGCGCGCGCGCTCCACCGGCGCGGC¹ 29400
E R L A A V I A G S G A A A V L H R P G

29401 CTCGAAGGGCGGCTGCCCGCGGGCGTCCGCGCGCTCCCA¹CCGACGCCCGCGACGGCAG²C 29460
L E G R L P A G V R A L P T D A A D G S

29461 ACCGCCACGCACGACCCCGGGCCACCCGCCACCCCGCAACCGCGCTACGTGATGTAC¹ 29520
T A T H D P G P T A T P R N A A Y V M Y

29521 ACCTCCGAGTCCACCGGAGAGCCCAAGGGCATCGTGTGCAACCGCAACGTCTGTGGCC¹ 29580
T S G S T G E P K G I V V E H R N V V A

29581 TCCCTGCGCGCCGCGGGGCCACTACGCGCGCGGACCGCGCGGTTCCTGCTGTGTCC¹ 29640
S L A A R G A H Y A A G P G R F L L A L S

29641 TCCTTCGCTTCGACAGCTCGGTGCGCGGCATCTTCTGGA¹CGCTGACCCAGGGCGGCACC² 29700
S F A F D S S V A G I F W T L T Q G G T

29701 CTCGTCTGCGCGGGGAGGGA¹CAGAACTGACCCCGCGCGCTGGTGGAGCATATGCC²L 29760
V L P G E G Q Q L D P A A L V S T I A

29761 CGGCAACGGCCACCCACCTCGCCATCCCTCCCTGCTGGCGCGCGTCTCTGGA¹CAG 29820
G Q R P T H T L A I P S L L A P V L D Q

29821 GCGCGCCCGCGGACCTCGCTCCCTGCGCA¹CGGTGATGCGCGCGGGGAGTCTGTGCG²A 29880
A A P G D L A S L R T V I A A G E S C P

29881 GCGGAAC¹TGGCCCGCGCTGCGCGGACCTGCTGCCCGGAGCACTTCCCA²CAACGAGTAC³A 29940
E L A A A C R D L L P G S T F H N E Y

29941 GGCCCCACGAGACCACCGTGTGGAGCACCGTCTGGTCCAGGAGA¹ACGACGACGACGGA² 30000
G P T E T T V W S T V W S Q E N E H D G

30001 CCCCACTCCCACTCGGCGCGCGGTGCGGGCACCTGGGTGCA¹CCCGCGGACCAACGC²C 30060
P H L P I G R P V A G T W V H P R D H R

30061 GGACGCA¹CGTCCCTCGGCGTGGCGGCGAACTCTCATCGGCGGCGCGCGCGTGGCC²G 30120
G R T V P L G V A G E L S I G G A G V A

30121 CGCGGCTACCTCGGGGCGCCCGGGACACCGCGCGCGCTTCGCGCCGACCCCGAGGCC¹ 30180
R G Y L G R P R D T A A A F R P D P E A

30181 ACGGCTCCCGGCGCGCGCTACGCCACCGGCGACCTCGGCGCTACCTCCCGGACGGC¹ 30240
T A P G G R A Y A T G D L G R Y L P A D G

30241 AACCTGGAGTTCCTCGGCCGCGCCACCA¹CAGGTCAAGATCCGCGGCTCCGGGTGAG²N 30300
L E F L G R A D H Q V K I R G F R V E

30301 CTCGGCGAGATCGAGGCGGTCTCGAACCCACCGAGCTCCAGCGGACCATCGTCATG¹ 30360
L G E I E A V L D T H P E L Q R T I V M

30361 GCACGCGGACCA¹CCCGGCGACCAAGTGTGTCGCTACGTCTCTCCGCGCCCGCGGCA²A 30420
R G D H P G D Q V L V A Y V L P A P G

30421 CGGCGGCGCAACCGCGACATCCAGGGGTACGTCCGCGACCGGTGCGCCGCTACATG¹ 30480
R R P E P A D I Q G Y V R D R L P R Y M

30481 GTGCCACCGCGGTGATGCTCTCGACGCGTACCGGTACCGCGCGCGCAAGGTGCAC¹ 30540

V P T A V I V L D A V P L T A A G K V D

30541 CGGGCCTCGCTCCCGCCCCAGCCACGCGCCAGCTCACCCTGGGACGAGTACGTGCGAG 30600
R A S L P A P S H A Q L L T R D Q E Y L R

30601 CCCGGCACCGACACCGAGCGGGCGCTCGCGCCATCTGGGCGGACGTCTCAAATGGAC 30660
P G T D T E R A L A A I W A D V L K L D

30661 CGGATCGGGGCGGTGACCGCTCTTCGACGTGCGGCGGAATCCCTGGCGCGCATGCAG 30720
R I S L A G D R F F D V G G E S L R A M Q

30721 GCCACCGCGCGGCCAACAGATGTTCCGACCCCGGTCTCGTCCGCGCGCTCTTCGAG 30780
A T C A A A N K M F R T R T R V S V R R L F E

30781 GCGCCCTCCCTGCGGGAGTTGCGCCACGAGATCGACAAGCCCGCCTCGCGGGCGCGGG 30840
A P S L R E F A H E I D K A R L A G G G

30841 ACCGGCCTCACCGGCCCGCGGCCCGCCACCGAGGTGCGCGCGAATGACCCCGG 30900
T G L T G P A A A P A T G G A A E *
M T P A (orf25)

30901 CGCGACACACCCACCCCGCTCTCGCGGCCACGCGCAGCATGTGTTCTGCACCGG 30960
C D T T H P L S P A Q R S M W F L A H R L

30961 TCGCGCCCGAGGTGCCCGCTTACACATCTGACCGCCATCGAGTCAACGGCACACCGC 31020
A P E V P A Y N I C T A I E L T G T P R

31021 GCGCGCGCGCGCTCGGGACGTGGTACGGCGGCTCGGCGCAGGACGAGGCGCTCGCA 31080
P A A L R D V V R R L G R R H E A L R T

31081 CGGTGTTCCCGTGGTGGGGAGACCCCGCCAAAGGGTCAACGACCGGGCGGCGCC 31140
V F P S V G E T P R Q R V T D R A A P L

31141 TCGGACCGTGGACCTCACCCACCTGACCCCGCGCGCGGAGGCGAGACCGCACGGA 31200
R T V D L T H L T P A A A E A E T A R T

31201 CGCTACGGTGGCGCGCGCGCGCGGTCGCGGCTCGACACCGGCCCCCTGGCGGAATGA 31260
L R C A A A R P F R L D T G F L A E W T

31261 CCCTGCTGCGCGCGCGCGCGCGCGCGCTGCTGCTCTCTCGTCCACACATGCTCT 31320
L L R R A P G H A L L V L S V H I V F

31321 TCGACGGCGGCTCGCTCCACGTGGTCTGCGCGAAGTGGAGGCGTACGAGACGCGCC 31380
D G G S L H V V C R E L E E A Y G A A L

31381 TCGCGCGCGCGCGGACCCCTCGGCACACCCCGCGCGGCTACGGAACGCGAGTGCAGGA 31440
A G R P D P L G T P A P G Y G R Q C R T

31441 CGCGGGCGCGGAACAGSACGAGGCCGGCGGGAGTTCTGGCGCGCGGAATGTCCGCGG 31500
R A A E Q D E A G R E F W R R E L S G A

31501 CGCACCCCGCAGACCGTCTTCGGGGCACCGGCGCGCGCGGACCGCGCGCGCGCG 31560
P P R T T V F R G T G R P A G P P A R A

31561 CCACCGTCCACTACGACACCGAGCATCGGCCCCGACCGCGGACTTCTGCGCGAGCACG 31620
T V H Y G T D D P A P T A D F C R E H A

31621 CGCTACCGGCTACGTGCTGCTGCTCGCGGCCCTCGCCTGCTGGTGGCGGTTACACCG 31680
V T G Y V L L L A A L A C L V A R Y T G

31681 GCGGACGACGAGTGGTGAATCGGCTCACCGTGGGACTCGCGGAGGACCCCGAAGGCGCT 31740
R T D V V I G S P V G L R E D P E G L A

31741 CCACCGTCCGCGCGATGCTCAACCTGCTGCGGCTGCGGCTCGCGGTCGACGCGACCCG 31800
T V G P M L N L L P L R L R L H G D P G

31801 GCTTCGGCGAGGTCTTGGGCCCGACCCGGGAGACGTGCTCGGCGCGCTGGAGCACCGCA 31860
F G E V L A R T R E T L L G A L E H R T

31861 CCACACCGTTTCGAGGACATCGTGACCGCGTGGGGCGCGACCGGACCGGACGTGAGCT 31920
T P F E D I V D A V G A D R D P D V S P

31921 CCCTCTTCAGATCTCTTGGCCACGAACGCCCGCGGCCACCGCGGTTCACGGCGG 31980
L F Q I L F A H E R P P A P P A L P G V

31981 TCCGTCGCCGCGCTGTAACCGTCCCGGCTCCGCGCGCCAAAGTACGAGCTCGCCGTCACCG 32040
R A R V V P V P A P A A K Y E L A V T A

32041 CCACCGGACGACGCCGAGCGGCTCGGCTGATCGTCGAGGCGGAGCACGGACACCGGGAACT 32100
T E T P D G L R L I V E A E H G H E P

32101 CGGCGGAACCTCGCCGCTTCGCGCGCCACTTCGGGCTCTGCTGCGCGCCGGGGTCCGGG 32160
A E L A A F A R H F G V L L A A G V R A

32161 CGCGGACACACCGCTGAGCGCGCTGCGGCTGCTCACGACGAGGAGCGCGCGCGCTCA 32220
P D T P L S R L P L L T D E E R R R L T

32221 CCGACACACGCGCCCGCGCACCGCGCGGAGGCCCTACCGCCCTCTGACCGGGCTGG 32280
D T T A P R T A P E A P Y R P L H R L V

32281 TCGAGGATCGCGCCCGCGCGCGCGCTGGCGGTCTGTCGCGCGCACGCGTCAACC 32340
E E S A A R R P D A L A V V G G T R H L

32341 TCAGCTACCGGAGCTGAACTCGCGCGCAACCGGGGTGCGCGCTGGCTGCGCGCGGCTG 32400
S Y R E L N C R A N R R A A W L R R A G

32401 GCATCGGCACCGAGGAGCTGGTCCGCGCTCCGGCTGGAACGCGGCCCGGAACCTCTCGTCT 32460
I G T E D V V G V R L E R G P E L L V S

32461 CGCTCTCTCGCGCTCAAGGCGCGCGCGCTACCTGCGCGTGCAGCGCGCGCTGCGCG 32520
L L A V L K A G A A Y L P V D P A L P A

32521 CCGAGCGGGTACGGCTGATGCTCGACGACGCGCGGCGCGCTGCTGCTCACTCGAGACG 32580
E R V R L M L D D A R A A L L T T E T A

32581 CGCTCGGCACCCGCGCGCGCGCGCGCGCTGACACACGCTGGAACGAGCCGCGCAC 32640
L G T P P A P A G T P V H H V D G P P P

32641 CGCGGACCGCGCGCGGGAAGACGCGGACCAACCGCCCGACCTGCCACGAGCTCG 32700
P T R P G D D A D H T G P D L P T S L A

32701 CCTACTCTCTACACCTTCGGGTGACGCGCGCGCGCAAGCGCTGGGCCCTCCAGCACG 32760
Y L L Y T S G S T G R P K A V A L Q H D

32761 ACAGCGCGCGCGGCTTCCTGCGCTGGCGCGCGCGCTTCGACGCGCGGAGCTGCGCG 32820
S A A A F L R W A G R A F D G G E L A A

32821 CGCTCTGCGCACCACTTCGCGCGGCTTCGACCTGTGCGTCTTCGAGCTGTTGCGCCCC 32880
V L A T T S A G F D L S V F E L F A P L

32881 TGGCCACGCGCGCACCGTCTGCTCGCGACAGCGCCCTGCAGCTGCCGCCCTGCCT 32940
A H G G T V V L A D S A L H V P A L P W

32941 GGGCGCGCGCGGACGCTCTGAACACCGTGCCCTCCGCGCGCGCGCGCTGCTGAGAC 33000
A P A A T L L N T V P S A A A A L L D A

33001 CCGACGCGCTGCCGACGGTCTGACGCGCGCTCAACTGCGGGCGAGCCCTGACCGCGG 33060
D G L P D G L T A V N L A G E P L T A E

33061 AGCTGTTGCGCGGCTGCACGCGCGCTGCGGAAGCGCGCTCGCAACCTCTACGCGC 33120
L V A R L H A R L P K A A V R N L Y G P

33121 CCTCGGAGGCGCACCACTACGCCACCGCGCGCTCTGTCGCCCGGGCGGACCGAGGCGC 33180
S E A T T Y A T A A L V P A G T E A P

33181 CGGCGATCGCGCGGCGCTCGGCGCGCGCGCTGTGGACCGCGACGACGCGGACGCGC 33240
A I G R A L G A A R V W T A D D R Q R P

33241 CCCTCCCGCGCGGTCTGCGTGAACCTCTCATCGCGGTACGCGCGCGCGCGCGCTG 33300
L P G A V V G E L L I G G T A P A R G Y

33301 ACCTCGCGCGCGGACGACCGCGCGACGCTTCGCGCCGATCCGACGGGACGCGCCG 33360
L G R P G P T A D A F R P D P T G P P G

33361 GCTCCCGCTCTACCGCACCGGGGACCTGGCCGTACGCGCGCGCGAGCGCGGTCTGCTG 33420
S R L Y R T G D L A V R R P D G R F V F

33421 TCCTCGCGCGCAAGGACGAGCAGATCAAACTCCGCGGGGTGCGATCGAACCGGCGAGG 33480

L G R K D E Q I K L R G V R I E P G E V
 33481 TGGAAGCCGCTCTCCGCGAGTGC GCGCGCGTTCGCGCGCGCGCGCTCGTGTCTCGCCGGGA 33540
 E A A L R Q C A P V A A A V L L A G T
 33541 CCACCCGCGAGAACACCGCTCGTCTCGCTTCGTACCCCTTCGCCCGCGCGCCCGCGTCTG 33600
 T A E N H R L V G F V T P S P G A R V D
 33601 ACCCCGAGCGCACCTCGCGCGCTGCGTTTCGCGCTTCGCCCGCGCGCCCTCGTGC CGCGCG 33660
 P E R T L A A L R S R L P A A L V P A A
 33661 CGCTGGTGTGCGACGCCCTGCGCTGACCGCGAAGACCGACCGCGCGCGCGCG 33720
 L V V C D A L P L T A N G K T D R A A L
 33721 TCGCCGCGCGCGCGCGGACACCGCGCGGACACCGCGCGTACGCCCGCGCGCGCACCC 33780
 A R R A R G H R P D H G A Y A P P R T R
 33781 CGCTCGAGAAGCGGTGCGCGCGATCTGCGCGAGGTGCTCGGGACCGAAACGGGTGGGGA 33840
 V E K A V A A I W R E V L G T E R V G I
 33841 TCCACCGAGGGTCTTCGACGCGCGCGCGCACTCCCTGTCTGCTGCGCTTCACCAAC 33900
 H Q G F F D A G G T S L S L L R L H H R
 33901 GGCTGGTGGCTCCGTCCATCCCGGCTCTCGGCTCGCGCGAGCTTTCGCTGCGCGACG 33960
 L V A S V H P G L R L A D V F R L P T V
 33961 TCGCCGCGCTCGCGCGTTCTGCGACCGCGAGGAGCGCGCGGACCGCGCGCGCGCG 34020
 A A L A A F V D G Q E D A R E T A V G D
 34021 ACGCGGCCCTCCGGGCGCGCGCGCGCGCGCGCGCGCGCGCGCGAGGAAAGCG 34080
 A A L R A G R R R R A A V A A R R R R K G G
 34081 GCGGACGATGAGCCATGCCGACGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCG 34140
 M S H A D A G D G L D A D T T G A C
 G R * (orf24)
 34141 GGCCGAGGATCGCCGTGATCTCGCTGGGCGGACGCTTCCCGGAGCGGACCGGGTGA 34200
 A D G I A V I S L G G R F P G A D R V D
 34201 CGCCCTCTGGACGAACCTGCTGACCGCGAGGACGCCATCAGCCACTTACCCGCGACGA 34260
 R L W T N L L D R E D A I S H F T A D E
 34261 ACGCTCGCCCGGGGCGCGGACCCGAACTGGTGCGCCACCGCGCGTTCTGCTGGCGCGGA 34320
 R L A R G R D P E L V R H P R F V G A E
 34321 AGGCGTCTCGGCGAGCTTCCCTCTTCGACGCGAGTTCCTTCGCGGTCTGCGCGCGGA 34380
 G V L G D V S L F D A E F G C S P R E
 34381 GGCCGAAGTCATGACCCGCGAGCACCGGCTCTGCTGGAGAGGCGTGGCACGCTCTTCGA 34440
 A E V M D P Q H R L C L E E A W H V F D
 34441 CACCGCGGCTACGACCCGCGCGCGGACCGCGCGCGCGCGCGCGCGCGCGCGCGAG 34500
 T A G Y D P A A T G T A V G V F L S A S
 34501 CCTCAGCTCGTACCTGATCCGCAAGTCTTCCCGCGCGCGCGCGCACAGCGCTCTGCTG 34560
 L S Y L I R N V L P G G A A Q R L L G
 34561 CGGCTTCCGCTGTGATCCACAACGACAAGGACTTTCGGCCACCAACCGTGTCCCAAA 34620
 G F P L L I H N D K D F L A T T V S H K
 34621 ACTGGGCTTACCGGCGGAGTTACGCGCTCGGCTCGGCTGCTGCTCTCCCTCGTCCG 34680
 L G L T G P S Y A V G S A C S S L V A
 34681 GGTGCACCTGCGCTGCCAGAGCTGCTCACCGAGGAATGCGCATGCGCGCTGCGCGCGG 34740
 V H L A C Q S L L T E E C D M A L A G G
 34741 GGTCTCGCTTCAAGTGCAGGCGGCGAGGTTACGTGACGCGGACGAGCGCATCTACTC 34800
 V S L Q V P Q G Q G Y V H A D D G I Y S
 34801 ACCCGCGGCGCTGCGCGCCCTTCGACGCGCGCGCGCGCGCGCGCGCGCGCGCG 34860
 P D G R C A P F D A G A A G T V G G S G
 34861 CGTGGGCTCTGCTGCTCAAGCGGCTCGCGGACGCGCTGCGCGAGCGGACCGCGTCCA 34920
 V G L V L L K R L A D A V R D G D R V H

34921 CGCGGTGATGTCGCGCTCGCGGTGAACACGACGGCGCGACAGGTGCGTTACACGGC 34980
A V I L G S A V N N D G A D K V G Y T A

34981 GCCCGCGTACACGGCAGAGCGCGTCTGCGCGAGGCGCTGGCGGTGGCGGGATCTC 35040
P G V T G Q S A V V A E A L A V A G I S

35041 CGCCGCGACCGTCTGCGGTCTTGAGGCGCACCGGCACCGGCTGGCGGATCCCGT 35100
A A T V G V L E A H G T G T R L G D F V

35101 CGAAGTGGCCGCGTACCCGGCGTTCCGCGCCACACGACCGGAGCGGCTTCTGCGC 35160
E V A A L T R A F R A H T D R S G F C A

35161 GCTGGGCTCGTGAAGGCCAAGTGGGCCACTGGACGCGGCGCGGCGTACCGGGCT 35220
L G S V K A N V G H L D A A A G V T G L

35221 GATCAAGGCCGTGCTGGCGGTCCGCGAGGGCGTCATCCCGGCACCCGCACTACCGTTC 35280
I K A V L A V R E G V I P G T P H Y R S

35281 GCCCAACCCCGCATCGACTTCGCCACACACCCCTTCTAGTCACCGCGACCCCTCGC 35340
P N P A I D F A T T P F Y V T A D T L A

35341 CTGGCGGAGCGGACACCCCGCGCGCGTCACTCTTCGCGATCGCGGCGAC 35400
W P E A D H P R R A G V S S F G I G G T

35401 CAACGCCCATGATCCTGGAACGGCCCGCGCGCGCGCGGACCGGACCGC 35460
N A H V I L E Q A P P A A P R A D R T A

35461 CGGGGTGCCCATGCCGTGTGTGTGTCGCGCCGACCCGCGAAGCACTGGCGAGGCGCT 35520
G V P M P L V V S A R T R E A L A E A V

35521 CGCGGACCTGGCGGCGTGTGTCGCGCCCGGAGCGCGGGACCCGGCTCGCGATCTCGCGC 35580
R D L A A W S A P E P G T R L A D L A A

35581 CACGTGCGCGCGCGCGCGCTTCCCGTACCGCGCGCGCTGCTGTGCGACGACCTGCC 35640
T L A G R R A F P Y R A A V V C H D L P

35641 CGAGCGCGCGCGCTGTCGCGCGCGCGCGCGGAGACCGCTCCCGCGCGAGGAGGC 35700
E A A R L L G G A R G E T A L P G R E A

35701 CGTGTCTCTCTCCCGCGCGAGGCAACCTCCCGCGGACACCGCGCGCGCTGTACGC 35760
V F L F P P G Q G T L P P D T G R G L Y A

35761 GGACGTGCCCGCGTTCGCGCCCACTTCGACGCGCTGTGCGAAGGGTTCGCCCCGCTCG 35820
D V P A F R A H F D A C A E G F A P L G

35821 CACCGACCTCCACGCGCGCTCGGGGCCCGGCGGACGACACGAGGCGCGCAACCCGC 35880
T D L H A A L G A P A D D T R A A Q P A

35881 CCTCTCGCCGTGAGTACGCCCTCGCGCGCACCTGATGACTGGGTGTGCGCGCGGCG 35940
L F A V E Y A L A R T L M D W G V R P A

35941 CGCGATGCTCGGCGACAGCTCGGCGAGTACGTGCGCGGACGCTGCGCGGGTCTGTC 36000
A M L G H S L G E Y V A A T L A G V L S

36001 CCTGCGGACCGCTGACGCTCGTCCGGGCCCGGCGGAAGCGACGACACCATGCGGCC 36060
L P D A L T L V R A R A E A Q H T M P P

36061 CGGCGCGATGCTCGCGTTCGCGCTCAAGCGGACGACCTGCGCGCGCTGCTGCCCCGGA 36120
G R M L A V P L T P D D L R P L L P L G

36121 GGTGAGTTCAGCGCTTCAACGCCCGCGCGCTGCGTCTGTCGCGGCGCGCGGAGCC 36180
V E F S A F N A P G R C V V G G P P E F

36181 GGTGGCGAGCTGCGCGCGCGTGGCGCGGCGGAGTGC CGGCGCGGACGCGGAC 36240
V A E L R A R L A R R G V P A A E L A T

36241 CGCGACGCGCTTCACTCGCGCGCGTGAACCGTGTGTCGCGGCTTCGCGGCTGCT 36300
A H A F H S A A V E P L L D G F R G V L

36301 GGAAGCGCTGCGACTGCGCGCGCGCGCTGCGGTACGTCTCTCCCTACCGGCGACTG 36360
E G V R L R P P R L R Y V S S L T G D W

36361 GCGCGACCGCGGTACACACCCCGCGTACTGGCTCGCCACCTGCGCGCGCGCGTCCG 36420

A D A A V T T P A Y W L A H L R R P V R

36421 CTTCGCGGACGGCCTGCGGCGCTGCCTGGACCTCGGCCCGCTCGCCCTGGTCGAGACCG 36480
F A D G L R R R C L D L G P V A L V E T G

36481 GCGCGGGGCGGACTGACCGGCTGGCCGCGCGCGCGGGCCCGGAGAGCCCTTA 36540
P R A G L T G L A R R A A G P G E P P Y

36541 CACCGTCGCGCTGCGGCGCCCGGACGAGGCGCTTCGCTGACCCAGCGGTCGCGCT 36600
T V R C L A A P D E A A S L T H A V A V

36601 ACTCTGGCGCGCTCGGGCTGCGCGTGGACTGGACGGGCTCCACGCGCCCGGGCGCCCG 36660
L W R S G G C A V D W T A F H R P G R P R

36661 CGCACCAACGTCGCCGCGCTACCCCTCCCAACGGGTACGGCACTGGATCGACGCGCGGA 36720
R T T V P G Y P P Q R V R H W I D A P D

36721 CGAGTCGGAACCCACGGAACCTCGCCACCGCCCTGCGCGGGAGTTGCGGACGCGCGGA 36780
E S E P T D L A T A L R A E L R T D G D

36781 TCOCGCGCTCGCGCTGATCAGCGGCGCGGACTGCGCACGGGGCTGAACCGGCTGTGCGC 36840
P P L A V D Q R P G L R T G L N R L C A

36841 CGCCCTGGCGCGGACTACCTGGCCACCGCGCTCGAAGCGAGCGGGTCTGCGCGGAT 36900
A L A R D Y L A T G V E A S G V L P G F

36901 CCACGCGCTTCTGGACTACCTGCGCACCTTGGCCGCTCGCGCACGGCGCGGACGACGCG 36960
H R F L D Y L R T L A A S A P A A D D A

36961 GGGGAGATCGCGCGGAGATCACGCGGCCACCGCTCTCTCGGGCTCGTCGACCT 37020
G T I A A E I T A A H P S F S G L V D L

37021 GCTCGGCGACTGCGCCAGGGCTATCGCGCGCCCTGTCACCCCGGAGCGCACTGGA 37080
L R H C A Q G Y P R A L S T P G A A L D

37081 CGTCCTCTATCCGCGCGGCGCGGCGACTCTCTGCGCGCACCTGGGCGAGGGCACCGC 37140
V L Y P A G S G D L L R R T L G E G T A

37141 CGACACCGCGCCACCGCGCGCTCACCGCGCTGGCGCGCTCCCTGCTCGACCGCGCTCGC 37200
D H R A T G R L T R L A G S L L D R L A

37201 GCGCGACCGGAACCGGCGCGCGCGCTGCGCGCTCTGAGGCGGAGCGGGCGCGGCGAG 37260
A D R E P G R P L R V L E A G A G A G S

37261 CCTCACCGGCGCGCTGCTCACCGGCGCGCGCGCTGCGACTACCAAGCCACCGACAT 37320
L T Q A L V T R A P G R L D Y H A T D I

37321 CTCCCGGCACTTCTGACCGCACTCGGCGGGAGGCGCGCGCGCGCTGACCTTCGT 37380
S R H F V T A L G R E A A R R G L D F V

37381 CGCGCGACCGCTCTCGACATCGCCCGGACCCAGGCGAAGGGCTTGGCGCGCGAGCG 37440
R A R V L D I A R D P G E Q G P A G E R

37441 GTTCGCGCTGCTGCGGCGCTCGCGCTGCTCCACGCGCACCCCGGACCTCGCGACCACT 37500
F D V V C G L D V V H A T P D L R T T L

37501 CGGCCATCTGCGCTCCCTGATGGCACCGGACCGCACCTCGCGCTGATCGAGACCAACGC 37560
G H L R S L M A P D G T L A L I E T A C

37561 CGACGACCCCTGGCTGACGATGATCTGGGGCTGACGGAAGGCTGGTGGCACCAACGA 37620
D D P W L T M I W G L T D G W W H H T D

37621 CGGGCGCACCCAGGCGCGCTGCTCGACGCGCGCGCTGGCGCGCGCTCTGGCGCGGA 37680
R R T H G P L L D A A G W R A L L A G E

37681 GGACTCTGCGCACGCGGATGTGATGTCGCGCGCGGACGCCCGGAGGCGCGCGCTGCT 37740
D F A T A D V I V P P D G P Q D A A L L

37741 GCTCGCGCGGAGACCCCGCGCGCGCGCGCGCACCGTCCGTGCGCAAGCGGAGCT 37800
L A R Q T P R P A A A A P S V G K R D V

37801 CGGCACGTGGTCTACGCGCGCGCTGCGCGGACGCGCGCGCGCGGACCCCGCGCT 37860
G T W C Y A R G W R H A A P A D P A P L

37861 GACGGGCGGCTGCTGCTGGGCGACGGGACACGGCGAAGCCGTCGCGAGCCGGCT 37920
T G G C L L L G D G D T A K A V A S R L

37921 GGAGGCCCTCGGCGTGCCCGTCAACACGTCGGCGGCGGCGACCGCGGGCCCCGAGCG 37980
E A L G V P V T T V G G G R P P G P E R

37981 GTACCGGGAACCTCGTCGGCCCCGCCACCGCTGCGCCGTGACCTGTGGCCGCTGCGCGA 38040
Y R E L V G P A T R L A V D L W P L R D

38041 CGCGTCCCACCGCGGCGCGCGCGCGCGCGCGTACGAGCCGCCAGGACGCGCGC 38100
A S H R G R A A G A A G V R T A Q D A A

38101 GCTGCACAACTGCTCCACCTCGCCCGGCGCTTCGGCGCGCTGGAGGAGCGCCACCCCGC 38160
L H N L L H L A R A P G A L E E R H P A

38161 CCGCGTGTGACCGTGACCAACCGGTGCCACGACGTGCTCGGCGACGACCTCGCCCAACC 38220
R V V T V T T G A H D V L G D D L A H C

38221 CGAGCAACCGACCGTCCCGCGCGGCCAAGGTGATCCCCCGGGAGTACCGCTGGATCGC 38280
E H A T V P A A A K V I P R E Y P W I A

38281 CTGACCGCCCTGGACGTGGAGCGGGCGCTGGAGCGCGAGCGGCTGGCGACCTGATCGT 38340
C T A L D V E P G L D A E R L A D L I V

38341 CCGGGAACTCGGCGCGCGCGCGAGACCAACGTCACCGCTCGCGCGGCGGACGCGCGCTT 38400
R E L G A A R E T T V T A C R G R R R F

38401 CACCCCTGCGCCGTCGGGAGCCCTCCCGCGCGCACCGGAACCGCGCGGTCCGCGC 38460
T P C P V R Q P L P A A P E R P A V R P

38461 CGGCGCGCTCTACCTCGTCTGCGGCGGCTCGGCGGCATCGGCCTCCACCTCGCCGAGTA 38520
G G V Y L V C G G L G G I G L H L A E Y

38521 CCTGGGCGGCGCGCGCACCAACGTCGTCTCACCCACGCGCGGCGCTTTCGCGCGCGG 38580
L G G A R T T V V L T H R R R P F P A P G

38581 CGCTGGGAGCGGCTGCGCGCGGACACCGGAGCGGCGGCTGCTCGGCGGCTGCGCTC 38640
A W D G L P A G H F E A A V V R R L R S

38641 CCTCGCGCCACCGCGCGCACGCTGCTGCTCGCGGCGCGACCTCACGACCAACGACGC 38700
L A A T G A T V V V R R A A D L T D H D A

38701 GATGCGCGCCTCGCGGACGAGGTGGAAACAGGCCACCGCGCCGTCGCGGGGTGGTGCA 38760
M R A L A D E V E Q A H G P V R G V V H

38761 CGCGCGGGGTGCCCGACACCGCGCGCATGATCCAGCGTCGCGACCGAGCCGCGACGGA 38820
A A G V P D T A G M I Q R R D R A G T F

38821 CGCGCCCTCGCGCGCAAACTGACCGGACCCCTCGTCTGGACGAGGTGTCGCGCCACCG 38880
A L L A A K L T G T L V L D E V F A H R

38881 CGAACCTGACCTTCTGTCCTGTCCTCGATCGGCGACCGTGTGTCACAACTGAAGTT 38940
D L D F L V L C S S I G T V L H K L K F

38941 CGGCGAGGTGCGCTACGTGGCGGGCAACGAGTTCTCGACCGCTATGCGCGCCACCGCGC 39000
G E V G Y V A G N E F L D A Y A A H R A

39001 GGC CGCGCGCGCGCAGAACCTGTGATGCGCTGGACCGACTGGCGGGAGTGGCGCAT 39060
A R R P G R T L S I A W T D W R E S G M

39061 GTGGCGCGCGCCAGCGCGCTGACCGAGCGCTACCGCACCGCGCGCGCTGCGCGCT 39120
W A A A Q R R L T E R Y G T G A D L F V

39121 ACCGCGCGGGCGGACCTGCTCGGCGCGATCAGCGCGGAGGGCGTGCAGCTCTTCGC 39180
P P G G D L L G A I S P E E G V D V F A

39181 CCGGCTGCTCGCGCGCACCGCGCGGACGTGTCATGCTGCGCGCGGACCTCGCGAGA 39240
R L L A A D T G P N V I V S A Q D L D E

39241 ACTCTCGCGCGGACGCGCGGTACACCGACGACCACTCGCGCGCGCTCGCGGACCT 39300
L L A R H A A Y T T D D H L A A L G D L

39301 GAGGATGCGCGCGCGGACCGCTCGCGCGCGCGCGCGTACGCGCGCGCGCGCAC 39360
R I A A A R D R S A P A A P Y A A P H T

39361 GCCGCCAGCGGGGATCGCGGCTGGTACCGGAGCTGCTCGCGCTCGAACACGCTCGG 39420
P A Q R R I A G W Y R D L L G V E H V G

39421 CCTGACGACGACTTCTTCCGCTCGCGGGGACTGCTGCTCGCCCTGCGCTGCTGTC 39480
L D D D F F A L G G D S L L A L R L L S

39481 GCAGCTGCGGAGCGCTACGGGGTGGAGATCTCCGTGCGCCGATGTTGACGAGCCAC 39540
Q L R D A Y G V E I S V A R M F D E F T

39541 GGTGGCGGCGCTGGCGCGCGCACGGCCGCGCGGAAGAGACGCGCGGCAGGAAGA 39600
V A A L A A A T G P P P E E T P G Q E E

39601 GGTGGTGTGTGACCACGCCCGCATCACCGACTGCTCAACGAGCTCGCGCGCGCGAG 39660
V V L *
M T T P R I T D L L T E L R G R Q (orf23)

39661 GTGACCTCAGCGCGACGGGGACGGCTGCACTGCCGCGCGCCCGGGGCGCGTCAAC 39720
V T L T A D G D R L H C R A P R G A L T

39721 GACGAGCTCTCGCCACCATCCGCGCCCGCGCGAGCACTCTCTGCGCCACCTGCGCGCC 39780
D E L L A T I R A R R D E L L A H L R A

39781 GACCGCGCATCCGCGCGCACGACGGCCGCGCGCTGCTCTCGCCCGCGCGCGCTC 39840
D R R I P R H D G P A P L S F A Q E R L

39841 TGGCTCTCCACAGTTCACCCGCGACGAGCGCTCAACACTCCCCCTGCACATCGCC 39900
W L L H Q P H P H D S A Y N I P L H I A

39901 CTGCGCGGCCCTGAACCCGCGCCCTGCGCGCGCCCTGGCGAGTGGTACGCGG 39960
L R G P L N P A A L R A A L A E V V R R

39961 CACGAGCTCTGCGCACCGGTAACCGCATCAGCGCGGCTGCCCGCGCGCTGCGAA 40020
H D V L R T R Y A I S R G L R P R V E

40021 CCGGCCACACGCGCGCTGCCCTGACCGACCTGACCGGGCTCCCCGCACACCAACGG 40080
P A H T P P L P L T D L T G L F A H H R

40081 GACCGCGAACTCCCGCGCTGCGCGCGACGAGGCGCGCCCTTCGACCTCGCCCGAG 40140
D A E L A R L A A Q E A R R P F D L A Q

40141 GGCCGGTGTGCGGCGCGGCTCTCCGAACGGCCCCCGAGGAGCACCGCTGTGTG 40200
G P V L R A R L L R T A P E E H R L L T G

40201 ACCGCCATCACATGCGCAGCGAGCTGGTGTGCTCGACATCTGCTCCGCGAATGGG 40260
T R H H I A S D G W S L D I L L R E L G

40261 ACGTTTACCGGGCAGGGCGGACGGCACACCGCGCGCTCGACGCCCTGCGCTGCGG 40320
T F Y R A G R D G T P A G L D A L P L R

40321 TACCGCGACTTCGCGCGTACGAGCGGAACAGGCGGAACGGCGGAGACGGCCGAGCGG 40380
Y A D F A A Y Q R E Q A E R P E T A E R

40381 TCGACCGTGGGACGGCACTGAGGGCGCCCGCGACACTCGACGTCTCGGCGCC 40440
S T R W A R H L R G A P A T L D V L G P

40441 CCGCCCGCGAACCTCCACGCGCGCGCGCGCACCGTACGAGCGGACCTTCCCGCGCC 40500
P P A E P S H A P A G T V R T D L P A A

40501 CTGCTCAGCGCGTGGGAGCTGGGGCGCGGCGCGCACCGCTCTTCCGCTCTCTG 40560
L V T G L R Q L G G R A R T T L F P L L

40561 CTGAGCGCTTGGGCTCGCCCTGGCGCGCGCGCGCGCGCTACGACGTGATGTTGGG 40620
L S A C C T G L A L A G P P G P Y D V M V G

40621 ATCCCGCTCGCGCGCGCGCGCACCGAACTGGAGCGCTCATCGCTGCTTCCGCGAC 40680
I P V A G R P R T E L E P L I G C F A T

40681 ATCGCGCGGATGGGCTGACGAGCGAGGACCGAGCGCTGACCGGCTCGCGCGCGG 40740
I A P M R L T S D G T E P L T R L A A R

40741 GCCCAGCAGCAGTCCAGGACGCGTGGAGCGGACCGAGCTCCCTTCGAGCGGCTCGT 40800
A Q Q H V Q D A L D G P D V P F E R L V

40801 CACGCGCTGCTCGGAGCGGGACCTCGCGGAGAACCCCTGTTCTCGGCGTGTTCGCC 40860
H A L R P E R D L A E N P L F S A S F A

40861 TTCCAGAAACACCCCGGACCGCGGTGCGCCTCCCGGCTTGACGCGAGGTGTGCGCC 40920
F Q N T P R T A V R L P G L D A E V L P

40921 TCGCCGCGGTGGCCCCAAGTTCGCGCTGGCCCTCACCGACGCGCGGGCGGACGCG 40980
S P P V A P K F P L A L T A T A R A D G

40981 GGAATGGGCTTGAGCTGGAGTTCGACCGGGACGGATCGCGAGCCGCTCGCGCGGG 41040
G M G L E L E F D R D R I A E P V A R G

41041 ATCTCTCAGCTCTTCCACGCGCCCTCGCGCGCGGTGCGCGACCCGAGGCCCGCGG 41100
I L T S F H A A L A R A V A D P E A P A

41101 GCGCCGCTACCGCGCGCGCTGGACCGGCGCGGGCGGAAGGACAGAGTGCCTCA 41160
A P V P A A A V D R R P G R E G H E C L

41161 CACGAGCCGTTGGCGGGCGCGGACGCGCCACCCGACCGCGTGCCTCAGCTGCGGC 41220
H E P V A R A A A R H P D A V A V S C G

41221 GGCACCCAGCTCAGCTACGGGGCGCTCGACACCCGCGCGAACGGCTGGCGCGGTGCTG 41280
G T Q L S Y G A L D T R A E R L A A V

41281 CGCGCCACCGCGCGCGCGCGGCTGGTGGCCCTGTGCTGCCACCGCGCGCGGAA 41340
R A H G A G P E R R L V A L C L P T G P E

41341 TGGTCTGTGCGCGCCCTCGCCATCCTCAAGTCCGGCGCGCCTACCTGCGCTCGACCC 41400
W V V G A L A I L K S G A A Y L P L D P

41401 GCGACCCCGCGGAGCGCGCGCTCGTTCGCGCGGACGCGGGAGCGACGCTGATCGTC 41460
G D P A E R R A S V A A D A I V A G T L I V

41461 TCCGACACCGCGCTTCCCGCTCCACCGCTCGACGTACCGGCCACCTCCCGGACGCG 41520
S D T A L P P L H R V D V T A T L P D G

41521 GCCCGCGAGCCACCGCGCGCGCTCTGCGCGGCAACCTGCTACGCGCTACAC 41580
A P E P T A R A V L P G N L A C Y A C T V Y T

41581 TCGGCTTCCGCGCGCGCGCGGAGGCGTGTCTGTCACCCATGCCAACGTACCGGGCTC 41640
S G S T G G P K G V L V T H A N V T G

41641 CTGGCGCGTGCCTGAGGCCCTGCCGCGCTGGACGCGCGCGGACCTGTGCGGAC 41700
L A A C R E A L P A L D A P R T W S A T

41701 CACTCGCGCGCTTGCATTTCTCGCTGCGGAGGTGCGGGCGCGTGAACCGCGGGA 41760
H S P C A F D F S V W E V W G P L T A G G

41761 CGCTCTGCTCTGTGCGCGGACGTGGCGCGCGCGCGGACGAACTGTGGGACACCTC 41820
R L V L V P P D V A R A P D E L W D T L

41821 CGCAGCAACAGGTGGAAGTCTCAGCCAGACCCCGAGCGCTTCCACACCTCTCGGCC 41880
R D E Q V E V L S Q T P S A F H H L L P

41881 ACCGCGTGC CGCGCGCGCGCCAGGCCACCGCGCTCGAACTCGTCTGTGGCGGCGAG 41940
T A V R R A A Q A T A L E L V L G G G

41941 GCGTGGAGCGCGCGCTCTGACGCGCTTGGTGGGACGCGCTGGGCGACCGCGCGCGGC 42000
A C E P A R L T P W W D A L G D R R P A

42001 GTGTCTCAACATGTACGGCATCACCAGAAACACCATCCAGCTCACCGTCCGCGCGGATGAC 42060
V N M Y G I T E N T I H V T V R R M T

42061 GCGCGGACCGGTGCGGCGAGTCCGCTCGGCGCGCGCTGCGCGGGCAGCGCGCGGACCTT 42120
A A D R S G S P V G R P L P G Q R A D L

42121 CTCGACCCCGCGCGCGCGCTCGCGCGGGCGGGCGAACTGTCTGCTCGCGCGGC 42180
L D P H G R P V A P G G R G E L F V G G

42181 GTCGGACTGGCGCGCGCTACCTCGGCGCGCGCGCTCACCGCGCGGAGCTTCTGCGG 42240
V G L A R G Y L G R P G L P T A R S F L P

42241 GACGACACCCCGGCTGGCGGGCGCGCGCGCTACCGCTCGGAGAGCTTGGCCGCGCTG 42300
D D T P G W P G A R R Y R S G D L A R L

42301 CTGCCGACGCGGGCCTGGACTACGCGGGCCGCTCGACGACAGGTCAAGGTCGCGGGC 42360
L P D G G L D Y A G R S D A Q V K V R G

42361 TACCGCGTCGAGCCGCGGAGACCGAAGCGCGCGCTGACCCATCCCGCGCGTGCGCCAC 42420
Y R V E P A E T E A A A L T H P A V R H

42421 TGCCTGGTCGTGCCACGCGCGACGGCGACCGGCCATCTCGCGCGTACGTCTGTCGCC 42480
C V V V P R G D G D R R H L A A Y V V A

42481 GACACCCGCGCTGCGACGGGCGCGGGCTCGCACCCACCTGGCGAGCGGCTGCCCGCGC 42540
D T R A C D G P G L R T H L A E R L P R

42541 CACCTGGTCGCGGCTCGGTGGTCTTCTGAAGCGGATCCCGCTGACCCGCAACGGCAAG 42600
H L V P A S V V F L K R I P L T R N G K

42601 CTCGACGTGGCGGCTTGCCCGACCCGGCGGCCACCGCGCACCGCGCCGGAACGCCCG 42660
L D V A A L P D P A A H R A P A R E R P

42661 CGCACCGCGACCGAAGGACCCCTCACCGGCTGCTCGCGCCCTCTGAAGCGCGCACCG 42720
R T A T E R T L T R L L A A L L K A P P

42721 GAGACCATCGGACGACGACCACTCTTTCGACCTGGCGCGGCGACTCCCTGACGGTCAAC 42780
E T I G T H D N L F D L G G D S L T V T

42781 CAGTTCACCTCCCGGTGGTGGAGGAGTTCCCGGTGGACCTCCCGTGGCGCGGGTCTAC 42840
Q F H S R V V E E F A V D L P V R R V Y

42841 CAGGCGCTCGACATCGCGACGCTCGCGTGACCGTGGACGACTTCCCGCGCGCGCGCGAA 42900
C L D I A T L A V T V D D F R R A E

42901 CGCACCGCGTACTGCGCGCCCTCGCGCGCGGAGGCGATGGAACCGCGGTCACGGCG 42960
R T A V L R A L A A A E A M E P G G T A

42961 GGGGAGTCCGCGGTAATCCGAGGAGTCCCGCGTACGCGCGGGGGCGCGCGCTCGCG 43020
G E S G N P E E S A A T A R G P A E

43021 GCGAAGCAACCCGCGCTCGCGCGGTGAGTCCGCGCGCGCGCGGTGGAGCCCGCGCTC 43080
A N E P G A A A R E S G A A P V E P A V

43081 CAGTACAGGAGTCCCGCGCTACGAAGGGGAGCCCGGACCGCAGCGAATGAACCTCGCG 43140
A V Q E S A A T K G E P G T A A N E L G

43141 GCTGAGGCACGGGAGCCCGGCAACCGCAGCGCAGGAACCGGACCGACCCCGCGCACCC 43200
A E A R E P G T A A Q E P G T D P R P P

43201 GCCGCCACACCGCAGGACCCCGCACACACCGCAGGAAGGACAGCGCGTGGCGGTGCC 43260
A A T P Q Q D P R T T P Q E G Q P C P R P

43261 GAATGAGCGGCGCGCGCATCGTCGACATCGCGCGCGCTCAGCGCGGACGACCCCGCG 43320
E M S R P A G I V D I A R R H A E R T P A (orf22)

43321 CCGTCCCGGCTACGCGTCTCTGCCCCGAGCGAGAGCGTCCGCTTCTCTTCG 43380
R P A Y A F L P D G E T E S V R F S F A

43381 CGGACATCGACCGGCGGCGCGCGTGGCGCGCTCTCCAGGACCGCGGCTGGCG 43440
D I D R R A R A V A A V L Q D R G L A G

43441 GGGAGCGGTCCTGGTCCCTATCCCTCCGGGCGCGAGTACGTCCAGGCGTCTCTGGCT 43500
E R V L V A Y P S G P E Y V Q A F L G C

43501 GCCTGTACGCGGCGTGGTCCGCTCCCTGCGACGAGCGCGCTCCGCGCGGAGCGCG 43560
L Y A G V V A V F C D E F R S G P S A E

43561 AACGGTCCGCGGATCCGCGCGGACGCCCGCGCGCTGCGCTGACCGCGCGCGCC 43620
R L A G I R A D A R P A L A L T A C A P

43621 CGGAGCGCGGCTCGCGCGCTGGCCACCTGGACGTGGCGCGCTCCCGACTCCGCG 43680
E A G L A G L A T L D V A G V P D S A A

43681 CCGGGCGCTGACCGACCCCGTCCGCGGACCGGACCGCTGCGCTTCTCTCAGTACACCT 43740
G A W T D P V A G P D A L A F L Q Y T S

43741 CCGGATCGACCCGCGCCCGCGGCGTCATGGTCGGCCACGGCAACTGCTGGCCAAAG 43800
G S T R R P R G V M V G H G N L L A N E

43801 AGCGCTGCATCGCCGCGCTCGGCGCACGACCGGGACTCCACCTTCGTGGGATGGCGCG 43860
R C I A A A C G H D R D S T F V G W A P

43861 CGTTCTTCCAGCATGGGCGTGGTCGCGCAACCTCTCTCAGCCCTCTACCTCGGGTCC 43920
F F H D M G L V A N L L Q P L Y L G S L

43921 TGTGGTGCTGATGCGCGCGATGGCTTCCTCCAGGCGCGCGCGCTGGTCGCGGCGCG 43980
S V L M P P M A F L Q R P A R W L R A V

43981 TCTCCGCTACCGGGCGCACACGCGCGGCGCCCAACTTCGCTACGACCTGTGTGTGCG 44040
S R Y R A H T S G G P N F A Y D L C V D

44041 ACCGGTTCGGCGGAGCAGACGGGCGGACTGGACCTGTCTGGGCTGGAAGGTGCCTACA 44100
R V G E D E R A G L D L S G W K V A Y N

44101 ACGGCGCGGAACCTGTACGGGCGGACACCTGCGACGGTTACCGACCGCTTCGCCCCC 44160
G A E P V R A D T L R R F T D R F A P H

44161 ACGGCTTCACCCCGCGCGCACTTCGCGACCTACGGGCTCGCGAGGCGACCTGCTGCT 44220
G F T P G A H F P T Y G L A E A T L L V

44221 TCGCACCGCGCCCAAGGAGTGGCGCGCCCGCACCTGACCGCGCGACCGCGCGCGCTG 44280
A T G P K G V P P R T L T A D R A A L R

44281 GCGCGGCGCGCTCGGCGCGCGCGGCGCGGCGCGGCTGGAACCTGGTCGGCAACG 44340
A G R L R P A G P G E A G L E L V G N G

44341 GCACCGCGCGCTCGACACCACTTCGGATCGTCGACCCCGACCGCGCGGGAGTGCC 44400
T A G L D T T L R I V D P A T A R E G P

44401 CGCCCGGAGAGGTTCGGCGAGGTCTGGGTGCGCGCGCGGCGTGGCACCGCGCTACTTCG 44460
P G E V G E V W V R G P G V A R G Y F G

44461 GCGCGCGCGGAGTCCGCGCGCTGCTCGCGCGCGCGCGCGCGCGCGCGCGCGCGCG 44520
R P P R E S A P L L A A R L P G G E G P Y

44521 ACCTGCGGACCGGACCTGGGCGCGCTCGACGACGGGAACTCTTCTACCGGACGCGC 44580
L R T G D L G A L H D G E L F L T G R H

44581 ACAAGGACCTCATCGTCTATCGCGGCGGAGAACCAACACCGCGACGACCTCGAAGCGG 44640
K D L I V I R G Q N H H P H D L E R T A

44641 CCGAGCAGGCGCCCGCGCGCTCGCGCGGACCTGCGCGCGCGCGGTTGCGGTGCCCGGG 44700
E G A H P A L R P T C A A A F A V P G D

44701 ACGGCGCGGAGCGGCTCGTCTCTGCTCTGCGAACTCACTCTCTACCGCGCGCTCGACCGG 44760
G A E R L V L V C E L T S Y R A V D P A

44761 CGCGCTCGCGGAGCGCGCTCGGCGCGCGCTCGCGCGCGCGCGCGCGCTCGCGCGGACA 44820
C A V A E A V R A A L A A R H G V A P H T

44821 CGCTGGTGTGCTGCGCGCGCGGCGCATCCCAAGACCAACGCGGGAAGGTGCGGCGCG 44880
L V V L R R G G I P K T T S G K V R R G

44881 GCCACTGCGGACGCGCTACCTCGACGGAACGCTCCCGTTTCAACGCGCGCTCGCGCTCC 44940
H C R T A Y L D G T L P V H T A V R L P

44941 CCGCGGGGAGGAGGCGACCGGAGCCCTTCCTCTGACCAAGACCGCGCGCTCGGCTGGCCA 45000
A G E E G T E A L P L T D T P G R L A T

45001 CGCGCTCGCGGACCTCGCGCGCGCGCGCGCGCGCGCGCGCGCTCGCGCGCGCGCTCG 45060
A L R D L A A A H A G L A G P L P G T D

45061 ACGAGCCGCTGAGCGCGCTCGGCTGGAAGTCTGCTCGCGCTCGCTCGCGCTCGCACACG 45120
E P V S A L G L D S L A S L R L H H H V

45121 TCCAGTCCCGCTACGCGGTGACCTGCGCGCTACCGCGCTGCTCGCGGACCACTTACC 45180
Q S A Y G V T L P V T A L L G D T Y R

45181 GCGGCTCGCGGAGTGACGCTCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCAAG 45240
R L A E L T L A A P R P A R A P E G Q V

45241 TCACCGGCGCTCTGGCGGCGCTTGACGACAGGCGAGCGCCCTGTGGTACGAAACAGGCGC 45300
T G V W R P L T H G Q R A L W Y E Q A L

45301 TCGCCCCGACGCGCGCGCTACCACTCGTCCGCGCGCTGGCCCTCCGCGGCCCGCTG 45360
A P H A A A Y H L V R A L A L R G P V D

45361 ACGAGGAGGCCCTCGCCGAGGCGGTCCGCGCGCTCGTCCGCGCACCCCGCCCTGCGGA 45420
E E A L A E A V R R V V R R H F A L R T

45421 CCCGCTTCGCGCTCCGCGACGCGGAAACGCGCGCGGACCGGACCGTACGGACCGGAGC 45480
R F A L R D G E P A R R T E P Y G P E L

45481 TGGACGTACGCGACGCCACCGGCTGCGCGGACCGGCTCGCGGAACACCTGGCGCGCG 45540
D V R D A T G L P A D R L R E H L A A A

45541 CGGCGGACCGCCCTTCGACCTGGCGCGCGGACAGGCCCGTGAGGCTGACGCTCTACC 45600
G D R P F D L A A G D R P V R L T L Y R

45601 GCACCGGCGCGGACCATCTCTGCTGCTGGTCCGCCAACCTGGTCCGCGACTTCTGGT 45660
T D G G H I L L L V A H H L V A D F W S

45661 CCCTCGTCTCTCTCTGGGCGACCTCCGCGCGGCCACGCGCGGCGAGGACCTGCCGCCG 45720
L V V L L G D L A R A H A G E D L P F A

45721 CGCCGAGGGGGACCCCGCGACGAGCGGACGCGACCGGACCGTACTGGCGGACCC 45780
P E G D P G D E A T D A D R T Y W R H R

45781 GGCTCGCGGACGCGCCACCGCGCTCGACCTGCCACCGACCTCCCCACCCCGCGGAGC 45840
L A D A P P A L D L P T D L P H P A E R

45841 GCGGCTTCGCGCGGCCACCCACGCTTCGCGCTGCCCGCGGACCTCACCGCGCGCTGA 45900
G F A G A T H A F R L P P D L T A R L T

45901 CGCGCTCTCCCGGGAACGCGCACTGCACTCTTACCACTCTCTCGCGGCCACCAAGC 45960
A L S R E R H C T L F T L L A H A G L

45961 TACTGCTCCACCGCTGACGCGGAGGACGACCTCGTCTGGGACACCTCTCTCGCGCGC 46020
L L H R L T G G Q D D L V V G T L L A R R

46021 GCGACACCGCGGAGCGCGCGCGCGCTCGGCTACCTGGTCAACCGCGCTCGCGCTGCGT 46080
D T A E A A G A V G Y L V N P L P L R S

46081 CGGTACGCGAGCGCGGGGAGACCTTACGGAACCTGCTCGCGCGCACCCGCGGACCGTGC 46140
V R E P G E T F T E L L R R T R T V L

46141 TGGACGCGGTGCGCACGCGGCCACCCCTTCGGGCGCGCTCTCTCCGCTCTCGCCCCCG 46200
D A V A H G R H P F G P L V S R L A P A

46201 CGCGCACCGCGCGCGCGCGCTCTGCGAGAGCTGTCTGCTCCAGCGGAGTACG 46260
R T F G R A P L L Q S L F V L Q R E Y G

46261 GCGACGAGCGGACGCGGTACCGCGCGCTCGCCCTGGGCGTGGCGCGCGCTGCGCGTGC 46320
D E A D G Y R A L A L G V G G R L R V G

46321 GCGGACTCGACCTGGAGGCACTCGCGTTGCCGCGCGCTGGTGCAGCTCGACCTCTCGC 46380
G L D L E A L A L P R R W S Q L D L S L

46381 TGAGCATGCGCGGCTCGGGGACGCGGTGACGCGGGTGTGGGAGTACCGCACCGCACTGT 46440
S M A R L G D G L T G V W E Y R T D L F

46441 TCACCGAGGCCCGGTCCGCGAGCTGAGCGAGGCGTTCGTCCACTGCTGCGCGCGCGCG 46500
T E A T V A E L S E A F V H L L R A A V

46501 TCGAGGACCGCGCGCGCGCGTGGAGACGCTGCCGCTCACCGCGCGCGGAGACCGGGC 46560
E D P G A P V E T L P L T G G R E T G P

46561 CGCGCGCGCGCGCTCGCGCGCGCGCGCGCTCCGCTGACCGGCTCGTGGCGCGCGG 46620
R R G P S A A R P A L P L H R L V A A A

46621 CGGCGCGCGCGATCCCGCACGACGCGGTGCTGCACTCGCCCCGCGACGCGACCGCCC 46680
A R R D P A R T A V V A L A P D G T A H

46681 ACCACATCAGCCAGGACCTGACCGCGCGCGGCCACCACTCGCGCGCGCGCTCCGCGC 46740

H I S H G A L H R A A T T L A A R L R R
 46741 GGGAGGCGCGCCCGGAGCGGCCCGTCTCGTCGAGCGGGGCGCCCTGGCTGC 46800
 E G A G P E R P V A V L V L V E R G P F H R
 46801 CGCTCGCTACCTCGGCATCTGACGCGCGGGGCCACCGTGTGCCCCCTGGACCCGGAGG 46860
 V A Y L G I L H A G A T V L P L D P E D
 46861 ACCCCCGGCAAGGCTCGCCCGGACGATCGCGAAGTCCGGGGCGGCGCTGTGCTTACCG 46920
 P P H R L A R T I A N S G A R L L L L T E
 46921 AGACCGGGACCGCTCGCGCGCGGCGAGGCGGCGGTCCCGGCGTACGCGCGCTGACCG 46980
 T G T A S R A A E A A G P G V R A L T V
 46981 TGCCTGAGGGTCCACCGGCGCGAGCGGTCTCGGCGGACGTCCACCCCGAGCAGTCCG 47040
 R E G A T G G E R F S A D V H P E Q S A
 47041 CGTACCTGCTGTACACCTCCGGTTCGACGCGGACCCCAAGGCGGTGCTGCTCCCGCACCC 47100
 Y L L Y T S G S T G D P K G V L V P H R
 47101 GGGCCATCGTCAACCGCCTCTGTGGATGCAAGAGACCTACCGGCTGCGCCCGGGGAGC 47160
 A I V N R L L W M Q E T Y R L R P G E R
 47161 GGGTCTGTGCAAGAGCGCGGTGACGTTGACGCTCTCGATGTGGAGCTGTGTGGCCGC 47220
 V L H K T P V T F D V S M W E L L W P L
 47221 TGACCGCGCGGGCGACCGTCTGTCATGGCCCGGCGGGACCCACCGGACCCCGCGCGAC 47280
 T A G A T V V M A R P G T H R D P A R L
 47281 TCGTCGCGGATCGCCCGGAGGCGGTACACCGTGCACTTCTGCTCCCTCGATGCTCA 47340
 V R R I A R E A V T T V H F V P S M L T
 47341 CCCCCTTCTCACCGAGCTCGCCCGCGGCAAGCGCGGCTGCCCGCGCTGCGGCGGTG 47400
 P F L T E L A R G T T R L P A L R R V
 47401 TGTGAGCGGGGAAGAGCTGCCCGCGCGCGGTGAACCGCGCGCGGACCTCTCGACG 47460
 C S G E E L P A A A V N R A A G L L D A
 47461 CCCGGCTGTACAACCTCTACGGCCCGACCGAAGCGCGCGTACGAGTCAACCGCTGCGCCT 47520
 R L Y N L Y G P T E A A V D V T A W P C
 47521 GCGCCCGCGGACCGGGGCGGTGCGGATCGGCTGCGCATCGCCAAACCAACCAACG 47580
 R P P E P G P V P I G L P I A N T T T E
 47581 AGGTCTCGACGCGCGGTGCGCCCGTCCCCCGCGGTGCCCGGAGCTGTACTG 47640
 V L D G R L R P L P R P V P G E L Y L G
 47641 GCGGCGCTGCTGCGCCCAAGCTACCAACGACCGGCGCTGACCGCGCGGCTTCC 47700
 G A C L A H G Y H H D F A L T A R A R F
 47701 TTCGCGCCCGCGGCGGCGCGCTACCGCAACCGGACCTGCTCGCCCAACCGGCGG 47760
 P A P G G G R R Y R T G D L V R Q R A D
 47761 ACGGGGCTGCTGTTCGCGGACGACGAGACGAGGTGAAGATCGCGGATCGGG 47820
 G A L V F R G R T D D Q V K I G G I R V
 47821 TCGAGCCCGGCGAGGTGCGGAGGCGCTTTCGGGCGCTGCGCGGCGTCCGACGCGCGG 47880
 E P G E V A E A L R A L P G V A D A A V
 47881 TCGTCCCGCACGACGCGGCGCTGCGGCGTACGCGGTGCGCGACCGCGTGGCGCGCC 47940
 V P H D G R L A A Y A V A D P V G P A P
 47941 CGCGCGGACCGCGTGGGAGCGGCTGCGGCGGCGGCGGCGGCGGCGGCGGCGGCGG 48000
 A A D A L R D A L R R R L P G H L V P A
 48001 CGCGCTTACCTGCTGAGCGGCTGCGGCTCACCCCGGCGGCGGCGGCGGCGGCGG 48060
 A L T L L D R L P L T P A G K L D R R A
 48061 CGCTGCCCAACCGTGGGCGGCGGCGGCGGCGGCGGCGGCGGCGGCGGCGGCGGCGG 48120
 L P H P S A P P P D G G R P P T T G T E
 48121 AACGGCTCGTGGCGGGTGTGGGCGGAACGCTCGGACGGGAAGTGTGCGGTGGACC 48180
 R L V A R V W A E R L G R E V V G V D R

48181 GGGACTTCTTCCCTGGGCGGCGACTCGTCCGGGCGCTCGGCGTGACGGCGGGCCCTGC 48240
D F F S L G G D S V R A L G V T A A L R

48241 GCGCCGCGGGGCTCCCGGTGACGGTCACGACCTCCTGCGCTGCCACCGTGGCGGCC 48300
A A G L P V T V T D L L R L P T V A A L

48301 TCGCCGCGCACGCGACGAGCGGGCGGATCGCGACGGCGCGACAGGAGACGCCCCCG 48360
A R H A D E R A D R R P A R Q E T P P G

48361 GCGCGTTCGCGCTCTGCCCGGAAGCGCGGCGTGC CGCGCTCGAGGAGCGCTACCGGA 48420
P F A L C P E A A G V P G L E D A Y P M

48421 TGTGATGGCCGACGGCGCGTGCTCTTCCACCGTGACACACCGCGCTACGAGGTCT 48480
S M A Q R A V L F H R D H N P G Y E V Y

48481 ACGTCACCAGCGTCGCGCTCTCCACGCCCTGGACCGCACACGGCTCGCGCGGCGGTG 48540
V T S V A V S T P L D R T R L A A A V D

48541 ACCGGCTGCTGGACGGCACGCTATCTGCGGTCTCTTCGACCTCGTGTCCACCGG 48600
R L L D R H A Y L R S S F D L V S H P E

48601 AGCCACCCAGCTCGTCTGGACCCACCTGCCACCCCGCTCGAGGTGGTGGAGTCTGCTG 48660
P T Q L V W T H L P T P L E V E S S D

48661 ACCCGCGCGTTTCGACGCGTGGCTGCACGCGGAACGCAAGCGCCCGCTCGACGTCGGCA 48720
P A G F D A W L H A E R K R P L D V G T

48721 CCGGACGCTGCGCGGTTTCCACGCGCACGACGCGGAGCGCGCGATTTCGCGGTGACG 48780
G P L A R F T A H D A G A A G F R L T V

48781 TCAGSAGCTTCGCGCTCGACGCGTGGTGGTGGCCACCGTGTCCACGAATGCTCCGCG 48840
S C F A L D G W C V A T V L T E L R D

48841 ACTACTGGTTCGCGCTGCGCGCGCGCCCTCAGCCTCCCGGCACCGCGCGCTCTACC 48900
Y W S A L R G A P L S L P A P A A S Y R

48901 GCGAGTTCGTCGCGCTCGAACGCGCGGCCAACAGATCCGCGCGCACCGGAGTTCGCG 48960
E F V A L E R A A Q H D P A H R E F W R

48961 GAGCGAGCTCGCGCGTGGCGCGCGCGTCCGCTGCCCGCGCGCGGTGCGACCGCGCG 49020
T E L A G A R P H P L P R R P V P P G

49021 GGCGGACGGGATCCGCGACGACCGTCACTGCTCCCGTGGAGGACCGTGGCAAGG 49080
P D G I R Q H R H V V P V E D T V A K G

49081 GCCTGCGCGCTCGCGCGGAGTGGGTGTCGGGCTCAAAACGTTCTGCTGGCGGTCC 49140
L S A L A G E L G V G L K H V L T G V H

49141 ACCTGCGGCTCGTCCGGGCGCTGTCGGGCGACCCGACGTCATCACGCGCGTGGAGACC 49200
L R V V R A L S G D P D V I T A V E T H

49201 ACGCGCGCTCGAACGGCACGACGGCGACCGGCTCTCGGGGTGTTCAACCAATCTGCG 49260
G R L E R H D G D R V L G V F N N I L P

49261 CGCTGCGCGCGGTGGACGCGCGGAGTGGGCGGACCTGGCCCGCGCGCGCGACGCGCG 49320
L R Q R V D G G S W A D L A R A A H A A

49321 CGGAGGCGCGACGGGGAGTACGCGCGCTATCCGCTGGCCGAGGCACAGCGACCAAG 49380
E A R T G E Y R R Y P L A Q A Q R D H G

49381 GCGCGCGCGGCTCTTCGACACCTCTCTCGTGTTCACCCACTTCCACTCTACCGCGCG 49440
A A G L F D T L F V F T H F H L Y R A L

49441 TGGCCGACCTGGACGATGGCGGTCTCCGACCTGCGGGCCCCCGACGACCTACGTAC 49500
A D L D G M A V S D L R A P D Q T Y V P

49501 CGCTCACGCGCCACTTCAACGTCGACGCCACGGACGGCGCGCGCTGCGGTGCTGTGG 49560
L T A H F N V D A T D G G G L R L L L E

49561 AGTCGGACCGCGGAGTTCGCCGACGACGAGTGGCGGAGTTCGCGCGGTACTACCGCG 49620
S D P R E F P D E Q V A E F A A Y Y R R

49621 GCGCGCTGCGGGCGCGCGCGACCGCGCGGCTACCGGGACACGCGGTGACGCG 49680
A L R A A A D A P H R P Y R D T P L T D

49681 ACCGGCCGGCCGGTCCGGCCCGCACCGCGCGGAGCGCTCCGTCCACGCCCTGTTCCGCG 49740
 R P A G P A P H R A E R S V H A L F A A
 49741 CCCCGGCCGGAACCAACCGGACCGGATCGCGCTCGACCGCGAGGACGGGCGCGTACAGC 49800
 P A R N H P D R I A L D G E D G P V S H
 49801 AC3GCGCCCTGCGCCGCGCGCCCGCGCTCGCCGGAACGTCGCGGCGCGCGGCGCGCG 49860
 G A L A R R A A R L A G T L R A A G A G
 49861 GGC CGGACACGTCGTGCGGATCTGGGCGCGCGCGCGCGGACGCGTCGTGCGCTGCG 49920
 P D T V V G I W A P R R A D A V V A L L
 49921 TGGCGCCCTCCACGCGGAGCGCCTACCTGCGCCTGGACCGGTCACCGCGCCCGCGCG 49980
 A L H A G A A Y L P L D P V H P P R R
 49981 GGCACGCGCAGGTGCTCACCGAGGCGCGCGCGCTGCTGCTGCTGCGCGCGCTCG 50040
 Q R Q V L T E A G A R L L V L P A G L D
 50041 ACACCCCGCTCCGGGCTCGGGCTGCCGTCGTGGCCCGGACGACCTCGGCGCGCCCA 50100
 T P L R A C G L P V V A P D D L G A P I
 50101 TCGCCCCCTGTCGTCCACCGGAGCACTGGCGCGGTCATGGCCACGTCGCGCTCCA 50160
 A P V S V H P E Q L A A V M A T S G S T
 50161 CGCGGACGCCCAAGACGATCGCGCTCCCGACGCGCGCCTGCGCGCTACCTCGCGTGG 50220
 G T P K T I G V P Q R A L A G Y L R W A
 50221 CGATCGGCTACCTCGCCTCGACGAGGAGACCGTCTCCCGGTGCACTCTCGTGGGCT 50280
 I G H Y R L D E E T V S P V H S S L G F
 50281 TCGACCTGACCGTCACCGCGCTGCTCGCACCGCTGGCGCGCGGCGCGGCGCGCTGA 50340
 D L T V T A L L A P L A A G G Q A R L T
 50341 CGGACTCGGCGACCGGGTTCCTCGCGCGGCACTGGCCGCGCGGACACACCTGC 50400
 D S G D P G A L G A A L A A G H H T L G
 50401 TCAAGATCACCCGCGCCATCTGGCGCGCTCGCCACAGTGGGCGCGCGGACCGCAC 50460
 K I T P A H L A A L A H Q L G A P T A L
 50461 TGCACCGCTCGTGGCGGGGGGAAACCCCTGCACGCGGCCACGTCGCGCCCTCGCG 50520
 R T V V A G G E P L H A G H V R A L R A
 50521 CCTTCGCGCCGGCGCCCGCTCGTCAACGAGTACGGGCGCGAGACACCGCTCGGT 50580
 F A P G A R L V N E Y G P T E T T V G C
 50581 GCTGTGCCACGACGTCGACCGGACCCCGGCGAGGCGCCCATCCCGTGGTACCCCGA 50640
 C A H D V A P D P G E A P I P V G T P I
 50641 TCGCGGCTCAGCGCGTGGTCTGTCGACGACGCGCTGCGCCGACCGCGCGGTGCGG 50700
 A G L S A C V V D D A L A P F G V R G
 50701 GCGAGCTGTACATCGCGGCGGCGGCTCACCGCGGCTACCTGGCGCGGCGCGGCCA 50760
 E L Y I G G T G V T R G Y L G R P A A T
 50761 CGCGCGCGCTACGTGCGGACCTGCGCGCCCGGCGCGCGCTACCGCACCGCGG 50820
 A A A Y V P D P A A P G A R R Y R T G D
 50821 ACCTGGCAGCGCGGCTGCGGACGCGACCCCTGCTCCTGGCGGGGCGCGGACCGCCAG 50880
 L A R R L P D G T L L L A G R A D R Q V
 50881 TGAAGATCGCGGCCACCGGTTGAACCGGGGAGGTGAGCAGGTGCTGCGCGCGCCACC 50940
 K I R G H R V E P G E V E Q V L G G H P
 50941 CGGGGTGCGGAGGCGCGGTCGTGCGCCACCGGACCGCGCGCGGCGCGCGCTGG 51000
 G V R E A A V V A H P A P G G G R R L V
 51001 TCGCGTACTGGGTACCGCGGAAACCGGCCCGGACCGTCCGCGGACGCTACCGCGC 51060
 A Y W V P A E P A R P P S A D A L T A L
 51061 TGCTCGCGACCGGCTGCGCGGTCACGCGTCCCGCGGAACTGTCGCGCTGCGCGCC 51120
 L A D R L P P Y A V P A E L V R L P A L
 51121 TGCCACCAACCCCAACGGCAAGGTGACGACACCCCGGCTGCCCGCGCGCGGACCG 51180

P T T P N G K V D H T R L P A A G R D R

51181 GGCGACTGGCGGAATGCTCGACCGGATCGAGGCATGTCCGACCGCGAGCGGCTCGG 51240
 R L A E L L D R I E A L S A G A

51241 CACTGCGCGACAGCGGCGCGACCCGGGAGTGGCGATGACCGAGCATGACGACCAACCGG 51300
 L R D S R P A P G S G D D R A *

51301 CGCGCCCGCGCGGCGCGCGCGGTTCGCTGGCGCCCGGCGGAAGCCGCGCCGTCGCGCAC 51360

51361 GTGCGGTGTCGCGGCGATGACGACCGGCTCGGACGCGCTGCGCGCGGACCGGAGCTGCCG 51420

51421 CGGACCGCGCGATTCTCTGGGGGACCCGCGGTTTCGGGTGGTGGCGCGCGCTCGTCCGCGAC 51480

51481 CCGGAGGTGCGCGATGCGCGGGCATGACGACCGCGCTCGGACGCGCTGTCCGCGGACTGGAGC 51540
 M R G H D D R V G R L S A D W S (orf21)

51541 GTCCCGCGGACCGCGCTGCGCGCGGGGACCGCGCGGTTTCGCTCGGCGCGCGGCGAGGC 51600
 V P P T R L P A G D P A G S V G P G G

51601 CGCGCGCTCCGCGACGAGGAGGTGACGATGTCGAGTATGACGACCGCTCGCGCGGCTG 51660
 P P V P H E E V T M S E Y D D R L A R L

51661 TCGGACACCAAGCGCGCCCTGCTGGACCGCTGGCTCGCGGAGGACCCCGCGCGGTCGC 51720
 S D N Q R A L L D R W L A E D P A G G A

51721 GGCCCGCTTCGCGCGCGCGCGCGCGCGCGCGCGGACCGGAGCGGATCCTGGCGCGGG 51780
 G P L R P D G R P P R T E A E R I L A G

51781 GTCTGGGAGGAGGTGCTGGAGACCGCGCGGATCGGCGCGGACGACGACTACTTCGCGCTC 51840
 V W E E V L E T G G I G A D D D Y F A L

51841 GGCGGAGACTCCGTCACCGCATCGTCATCGTGGCGAAGCGCGCGGACGCGCGGATCGCC 51900
 G D S V H A I V I V A K A R Q A L G

51901 CTGACCGCCCATGACCTCTTCGAGGCGAGGACCTCGCGCGCTGGCGCGGAGAGCGCC 51960
 L T A H D L F E A R T L A A V A R R A A

51961 CGCGCGCGCGCGCGCGGCGCGCTCCCGACGCGCGCGCGCGCGGTCGCGTACCGCTG 52020
 P A G P A E P V P D A G G G A V R Y P L

52021 ACCCTTATGCGAGCGGCGATGCTCTACCACTCGGCGCGCGCGGACGCGCGCGCTAC 52080
 T P M Q Q G M L Y H S A G G S T P G A Y

52081 GTGTGCGAGGTGTGCTGCGCGGCTGACGGGGGACCTCGACCTGGCGCGCTTCGCGACGCC 52140
 V V Q V C C R L T G D L D V A A F R T A

52141 TGGGAGCGCGTGTGCTGCGCAACCGCGCGTGGCGGTCTCTTCCACTGCTCGAGCGG 52200
 W Q A V L S A N P A L A V S F H W S D G

52201 TCCCGCGCGGAGCGGTGTGTGACCCCGACGCGCGCTCACCGTGGACACGCGCGAGTGG 52260
 S P P E Q V V B P D A R V T V D T A D W

52261 CGGACCGCGCACCGCGCGGAGCGGACGATGCTTCGCGCGCTTCGCGACACGACCGC 52320
 R D R T P A E R D D A F A R F L D T D R

52321 GCGCGGGCTTCGACCTCGCGCGCGCGCGCTGATGCGGCTGACGCTCTTCGCGAGGGC 52380
 A A G F D L A R A P L M R L T L F R G A

52381 GAGCACGCGTACCGCTGCGTGTGGACCCACCACTCGTCTCGACGCGTGTGCCAG 52440
 E H A Y R C V W T H H H L V L D G W S Q

52441 CAGCTCGTCTGCGGACGCTCTGACTGCTACATGCGCTCGCGCGCGGACGCGCGCGC 52500
 Q L V L R D V L D C Y M R L R A G R G A

52501 GAGCGCGCGCGCGCGCGCTCTTACCGGTATCTGCGCGCGCTGGAGCGGACGAGGCGG 52560
 E P P A R P S F T G H L R R L E R Q D G

52561 ATCGACGAGGAGTCTTGGCGGACCACTCGGCGGCTGCGCGCACCTCCCGCGTGGCC 52620
 I D E E F W R D H L G G L P A P S R V A

52621 GGTCCGCGCTGCGCGGACGCGCGGTTGGTCCGCGTACGCGCGCGGAGCACCGGACCGG 52680
 G P G C R D G R V V A V R A E H R H R

52681 GTCTCCGCGCGACGGGCGGGAGCTGACCGGCTTCTGCCGCGCCACGGGCTGACCCCG 52740
V S A A T G R E L T G F C R R H G L T P

52741 GCGCCGCTGCTGACGGGCGGTGGGCGGTGCTGCTGCTGCTGCACTGCGCGCAGGACGAC 52800
A A V L H G G W A V L L S L H C G Q D D

52801 GTGGTCTTCGGCACCACTCTCTCCGGCGCCCCGAGGACCTGCCCGCGTACCCGAGTGC 52860
V V F G T T L S G R P E D L P G V T E C

52861 GTCGGCTCTTCATCAACACGCTTCCCTGCGGGTCCGTTGCGGGAGGACACGGACGCTC 52920
V G L F I N T L P L R V R C G E D T D V

52921 GTCGACTGGCTCCACGGCGTCCAAAGCGACTTGGCGCCCTGTGGACACGCGCACTGC 52980
V D W L H G V Q S D L A A L W D H A H V

52981 CCGCTCAGCCGCGTcGAGCGCGGTCTCGGACTGGGCGGGCGGGCGGGCTGTCGACAGC 53040
P L S R V E R G L G L G R G G G L F D S

53041 ATCATGGTGTGCGAGAACTTCCCGCGCGCTGCGCGACCGGACGAGGCGGGCGGGCTG 53100
I M V V E N F P A A V A D G H E A G G L

53101 CGGTCGACGAGCCCCGGGCACTCGTGCACGAGGGCTACCCCTCGTACTGGAGGCCACC 53160
R V T E P R A L V D E G Y F L V L E A T

53161 ACCGGGACCGCGCGTGTCTGCACGCCCGCTACGACCCCGACCCGCTCGCGCGCGGGCG 53220
T G D R P V L H A R Y D P H R L A G G R

53221 GTCCAGGCGCTGTCTGCCGCTTCGACGACTACTCCGGGCGGTGACCGCGACCGCGCC 53280
V Q A L L A A F D D Y L R A V T A D P A

53281 CGCCCGCTGCGCGGACCTCGCGCGGTCTGGCCCGGACACCGCGCGCGGGACGCGCG 53340
R P L P D L R A V L A R D H A R R D A

53341 GCAACGGGCGCGCGCGCGCGCGGACCGCACCGCTGTGACGCTGGCCCGCGCGCGCGCG 53400
A R G R R R R A A D R T R L T L A R R R P

53401 GCGACGACGACCGAGGAGACCGCTGACATGACCGTGTGACCGGAGCGCGCGCGCT 53460
A T T T E G E T P * M T W T V T G A G G F (orf20)

53461 TCATCGGCTCCACCTCGTACGCGCGCTCGTCCGGGACCGGACCGGGTCCGCGGCTGG 53520
I G S H L V R R L V R D G H R V R G V D

53521 ACGTCTGCGCGCGCTACGCGCCCCGCGAGGCCACAGGAGTTGTCATCGCGGACCTGC 53580
L V P P R Y G P G E A Q E F V I A D L R

53581 GCGACGCGCGCGAGGCGCGCGGGCGTTCGCGCGCGGACTCCGCTCTTCGCGCTCGCG 53640
D A A Q A A R A V A G A D S V F A L A A

53641 CCAACATGGGAGGCATCGGCTGGACCCACACCGCGCGCGGAGATCTCCACGACAAAC 53700
N M G G I G W T H T A F A E I L H D N L

53701 TGCTGATCTCCACCCACACCATCGAGGCATGCGGGCGCGCGGCGTGCACACCGCTCT 53760
L I S T H T I E A C R A A G V R T T V Y

53761 ACACCTCTCTCGGCTGCGTCTACCCCGCGTCCCTGACGCGAGCCGACCGCGCGCGCG 53820
T S S A C V Y F A S L Q R E P D A A P L

53821 TGGCCGAGGACCGCGTCTTCCCGCGGAACCCGACATGGAGTACGCGTGGGAGAAGCTGA 53880
A E D P V F P A E P D M E Y G W E K L T

53881 CCACGGAATCTGTGCGGCGCTACCGCGGACGACGCGCATGGACATCAAGACAGCCC 53940
T E I L C G A Y R R S H G M D I K T A R

53941 GGCTGACCGCATACGCGCCCCCTCGGACGTCACACGGGCCCCGCGGAGTCCCTGT 54000
L H A I Y G P L G T Y T G P R A K S L S

54001 CGATGCTCTGCGACAAGTTCGCGGATACCGGCGACGAGGGGAGATAGAGGTCTGGG 54060
M L C D K V A R I P G D E G E I E V W G

54061 GGGACGGGACGAGACCGCTCCTACTGTACGTGACGACTGTGTGAGAGGCTGATCC 54120
D G T Q T R S Y C Y V D D C V E G L I R

54121 GGCTCGCCGCTCCGACGTGGGGAACCGGTCAACATCGGCTCCGAGGAGCGCTCGACA 54180

L A R S D V A E P V N I G S E E R V D I
 54181 TCGCGTCGCTCGTCGAGCGGATCGCCGGGTCGCCGGGAAGAAGTGCCTGCGCCTTCG 54240
 A S L V E R I A G V A G A F A
 54241 CCCCCACCGCCCGGTCGGCCCCCGGGCGCGTCTCGGACAAACCCGCTGCCGGAAC 54300
 P D R P V G P R G R V S D N T R C R E L
 54301 TGCTCGGCTGGSCACGGAGACGTCCTCGCGGCCGCTGAGCGACCTACCGTGA 54360
 L G W A F E T S L A A G L E R T Y P W I
 54361 TCGAGCCCGAGTCTCGCCGAGGCCGGGCGGATGCTGAGCACCGCACCGGTG 54420
 E R Q V L A E A G R A D A * M (orf19)
 54421 AAGGACCTCGGCCGCTGCTGCTCGGCCACCGCGCGCTTCCGGGGCCGCGAGCTGCAG 54480
 K D L G R L L L L G H A A R F R G R C A F
 54481 GACGTCGCCACCCGCGGCTGCGGGCCTCGCGGGGAGAAAGCTGGTGGTTCCTGTC 54540
 D V A T R A L R A S G G E N A W V V S V
 54541 GTCAACACAGTCTCCGGCCCGCAGGCCGCTGGACCAACCGCTGCGGCTCGCCCCCGC 54600
 V N T S L R A R Q A V D H A L R L A P R
 54601 CGCGGGCTTCCCGGCTGCGCTACCGCTTCTCCGCGGCCACACACGGCCACCCCGCC 54660
 R G L S R L R Y P P F S A A H H T A T P P
 54661 CGGACCTGTGCTGCTGCTGCCGACCCGGAACGCTCGGCAACGTCGAACGCTTCTC 54720
 R T L S L L C P T R E R V G N V E R F L
 54721 GACAGCTGCGCCGACCGCCCGCGGCCGCGGATAGAGGCCCTCTTCTACGTGAC 54780
 D S I G R A R T A A A P G R I E A L P Y V D
 54781 GACGACGACCCCCAACTCCCTGCCTACCAAGAGTGTTCGAGCACCGCCGCTGGCGTAC 54840
 D D D P Q L P A Y H E L F E H A R W R Y
 54841 GAGCGGATCGCGCGTGCSCCTGCAGTCGGCGCCCGCTCGGCGTACCCGACGCTGG 54900
 G R I G R C A L H V G A P V G V F H A W
 54901 AACCACTGGCCGGAACCGCGGCCGGAAGTGTGATGATGGCAACGAGACGACGTC 54960
 N H L A R N A A G D V L M M A N D D Q L
 54961 TACATGACTACGGCTGGGACACCGCCCTCGACGCCCGGTACCGAACTGAGCCGCCG 55020
 Y I D Y G W D T A L D A R V T E L S A L
 55021 CACCCGACGCGCTCCTGTGCTGTACTTCGACGAGCGCCAGTACCCCGAGGGCGGCTG 55080
 H P D G V L C L Y F D D G Q Y P E G G C
 55081 GACTTCCCGATGTTGACACGGCCCTGGTACCGCACCTCGGCTACTTCAACCCGACGATC 55140
 D F P M V T R P W Y G T L G Y F T P T I
 55141 TTCCAGCATGGAGGTCGAGAAGTGGTCTTCGACATCGCGACCGGCTGCACCGGCTC 55200
 F Q Q W E V E K W V F D I A D R L H R L
 55201 TACCCGCTCCCGGCGTCTCTGTCGAACACCGGCACTACCAAGGACTACCAAGGACCCCT 55260
 Y P V P G V L V E H R H Y Q D Y K A F
 55261 GACGCCACTACCAAGCGCACCGGATGACACGGGAGAAGTCTTCGCGGACCAACCGCCCTG 55320
 D A T Y Q R H R M T R E K S F A D H A L
 55321 TTCTTGCGCAACCGGCGGACCGGAGGCGGAGACGAGGCTGCGGGCCGTCATCGCC 55380
 F L R T E P D R E A E T D R L R A V I A
 55381 CGGGCAGGGAACACCCCGGACCGGACCAACCGGACCGTTCACGACCGGAGAC 55440
 R A G N T P D A D H A D H A V H D A E T
 55441 TTCTGGTTCACCGGCTCTGCGCGAGTCCACGCGAAGCTGCTCGCGGAACCTGACGAC 55500
 F W F T G L L R E S H A K L L A E L D D
 55501 GCGCCGGCCCGCGCGCGAGCGGTGCTCTTCGCGGACGGCTCTGGACCGCGCTGCC 55560
 A P G P A A G A V L F A D G V A
 55561 TACGCAACCAACCGCTGGCCACCGCCCTGCTCGCTCGATCCCGAGGCCACCTCGAC 55620
 Y R T H P L A T A L L A S I P E A T L D

55621 TCCGGCCGCGCCGACCTCTCGTCTCCGCGCGCGCTCCACCAACACCCGACGGC 55680
S G R A D L L V V P P G A S H H H P D G

55681 ACCGTCGACTCCGCGTTTCGGCTCCGACGCGCGCTCCGCGTCTGTTCCGACTCGCGGTG 55740
T V D S A F G S D A G L R V L F G L R V

55741 CCGGACGCGCGCACTCCGCTCGGCGACGCGCCCGTGCCTTGGGCAATGGGCAATGC 55800
P D A A Q L R V G D G P V P W G N G Q C

55801 CTGATCCAGCACACCGCGCACCGAGCACCTTGGCAACGACGCGACCGAATCTCTGGCC 55860
L I H D T A A P S T L R N D G T E S L A

55861 GCCCTCACCTTCGTGGTCCGCGCCCGGACCGGGGGAGTGAGGCCCGCTGTGCGGCATCG 55920
A L T F V V P R P A P G E *
M R P V C G I V (orf18)

55921 TGGCGATCCGCTCCGCGACGCGGGAAGTCTGACGCGCGTGAAGTACCGCGCCGATGGCCG 55980
A I R S A D G G L D G G E L T A P M A D

55981 ACCTGCGCGCGCGCGCCCGACGCGCAAGGCACCTGGGTCTCGCCCAACCGCGCGGGCGG 56040
L R P R G P D G E G T W V S P T G R A A

56041 CCCTCGGCCACACCGCGCTCGCGTGTATGCCCCGCAACCGGACGCCGCGGTGCGCC 56100
L G H T R L A V I A P D A G R Q P V A G

56101 GCCCGACGGCACCGTCCGCGTCTGTCGTAACGGCGAGTCTACGCGCTACCGGGAGATCC 56160
P D G T V R L V V N G E F Y G Y R E I R

56161 GCGGGAAGTCCGCGCGCGCGGCTCGCGTTCGCGCACCGCGACGACACGAGATCGCCC 56220
A E L R A A G C R F R T G S D S E I A L

56221 TCCACCTGTACTCGGGGACGCGCGCGCGCACTGGAGCGGCTGCGCGCGAGTTCGCGCT 56280
H L Y L R D G R R A L E R L R G E F A F

56281 TCGTCTCTGGGACGACGCGCGCCACCTCTTCCGCGCGCGCGACCGGTTCGGCGTCA 56340
V L W D E R R A T L F A A R D R F G V K

56341 AACCCCTCTACTACCGAGCGCGACGCGCGCTCTACGTCCGCTCGACGCTCAGGCGCT 56400
P L Y Y T E R D G R L Y V A S T V R A L

56401 TGCTCTCTCGCGCGCGCGCGCGCTGGGACACCGCGCTTCCGCGCGCACTGACG 56460
L S C G A P A R W D T A A F A A H L Q L

56461 TCGCGCTCGCCCCGACCGCACCTCTTCCGCGCATCCGCGAGCTCCGCGCGGCTCGC 56520
G L P P D R T L F A G I R Q L P P G C H

56521 ACCTCATCGCGACGCCACCGCACCGCGTCAACCCCTACTGGGACCTGACTACCCG 56580
L I A D A H G T R V T P Y W D L D Y P P

56581 CCGCGCGCGAAGTCCGCGCGCGGGAAGCTGGACGACCACTGGACGCGGTACGCGAAC 56640
A G E L A A R G S L D H L D A V R E R

56641 GGACCGACGAGGCGGTACGGTTGCGTACCGTCGCGACGTGCCCTCGCCTGCCACTCA 56700
T D E A V R L R T V A D V P L A C H L S

56701 GCGGGGCTGAGACTCTCCGCGGTGCGCGCTCCGCGCGCGCGCACACCGGCTCAGG 56760
G G L D S S A V A A S A A R H T R L T A

56761 CTTTACCGTCCGCTTTCGACGACCCCGCTTCGACGAGAGCGCGTTCGCGCGGCGCACCG 56820
F T V R F D D P A F D E S A V A R R T A

56821 CCGCCCACTGGCCATCGACCAACCGGAAGTTCGCTCGGAAACGCGCCCACTTCGCGGAC 56880
A H L A I D H R E V A S E R A H F A D H

56881 ACCTGCGGACGTCTGCGCGCGCGGAGTGGTGACGAGAACTCGCACGGCTACGCC 56940
L R D V V R A G E M V Q E N S H G I A R

56941 GGTACTGACAGCGCGCACATCAAGAAGCGGGATTACCGCGCTCTCGCGCGGGAGG 57000
Y L H S A H I K K A G F T A V L A G E G

57001 GCGGGGACGAAGTCTCTCGGCTACCCCAAGTTCGCGAAGGACCTGACGCTCAGCCTGT 57060
G D E L F L G Y P Q F R K D L T L S L S

57061 CCGCCGACGCGCGACAAAGCCGACCGCGGCTACGCCCGGCTGGTCGCGGCCGGGCTCC 57120
A D A R D K A D R G Y A R L V A A G L L

57121 TGCCGCGCTACCTGCGCACCTCTCTCGGCACCTCGGCTTCTCGCTCCCTGGATCGTCG 57180
P P Y L R T L L G T L G F L P S W I V D

57181 ACCGCGACCTGCGCGCTCACCCAGCCCGTGGCGCGCTGCTCGCGCCGACTTGGCGCGCG 57240
R H L A V T Q P V A L R P D F A A A E

57241 AACTGGCCCGCGCGACGCGCGCGCGCTGCTCGCGCGCGCGCGCGCTGCTCGCGCG 57300
L A R A D A A A P L L A A G A G L L A G

57301 GGCGCGCCCGCGCGCACCGGCCACCTACTCTTCGCCAAGACCTGGCTGCCCGGCTACC 57360
R A P A H Q A T Y L F A K T W L P G Y L

57361 TGCTCGCGCGCAACGCTCGACGCGGCCAGGCGCTGAGGTCGGCGTGCCTCTCTTCG 57420
L A A E R L D A A Q A V E V R L P L F D

57421 ACCACCACTCTTCGACCTCGTCGCGCACACCCCGCGGCTGGTACGACAAAGACGCGCA 57480
H H L F D L V R H T P P A W Y D K D G T

57481 CGCGCAAGTACCGCTGCGGCGCGCCATGCGCCACCGGCTGCGCGGAGGTGACCGAGG 57540
G K Y P L R A A M R H R L P R E V T E G

57541 GCGCGAAACAGGCTTCTCGCACCTCGATGGCGGACGACACCTCTCTCGACGCC 57600
P K Q G F L A P P M A D D D T L L D A C C

57601 TGCGGCAAGCCTCGCGCGACCGGCGCGCGGCGACGACCTCTCTTCGACCCGACGCGCG 57660
R E R L A G P G A G D D P F F D P H A V

57661 TCCGCGCCTGCTGGACCGGCTGGCGCGCCACCCCGCGGCGAGCGTCCGCGCGCGAGA 57720
L A L L D R L A A A P P G Q R S G G E K

57721 AACTCTCTCACTCGTGGAGCACCGCGAAGTGGCGAGAGTTCGGCTCACCACCG 57780
L L Q L V A S T A E L A D E F G L T T A

57781 CCCCGCGCGCGAGAAAGGCGGCAACGCTGCTGACCTCGATCCCGGCAAGCTCTCCGAG 57840
P S G Q K G G N G G

57841 GCCGAGCTGACCGCGCGATCGCGCCCTGTCCCCCGAAGCGCGGCGCGCTTCGAGAAG 57900

57901 ATGCTGCACGCGCGCGCACCCCGCGCCCGGCGATCCGCGCGCGCGCGCAACCGCGGCA 57960
M L H G A A A H P R P G I P R R G A T A A (orf17)

57961 CGGCTCTCTACGGCGAGGAACGCTGTGGCTGCTACCGGCTGCTGCCACCGCCTAC 58020
P A S Y G Q E R L W L L T G L L P T A Y

58021 AACTACGCCACCGCCTGCGGCTGCGCGCGACCTGTCCGTCGCCGCGCTGCGCGCGCC 58080
N Y A T A L R L R G D L S V P A L R G A

58081 CTGCGCGGATCTCTCGCGCGCACGAGTGTCTGCGCACCTCTCGGCTGGACGCGGAC 58140
L R G I V R R H E V L R T T F R L D G D

58141 GAGCTCATCCAGGTCTCCACCCACGGCGGACGTCCTCCGTCGCGCTGGCGGACCTAC 58200
D L I Q V V H P T A D V P V R L A D L T

58201 GGAGCTCTCCGCGACACCGGCGGCTGATGCGCGAGGAGGCGCGCGCCCTTCGACCTG 58260
G R S A D T G R L M R E E A R R P D F

58261 GAGCACGCGCGCTGCTGCGGCTGACCTCTTCCGCTCGGCGCCCGCGACCACTCGCC 58320
E H G P L L R L T L F R L G P R D H L A

58321 CTGCTGGCGTCCACCAACCGCTCACCGACGGCTGGTCCAAACGCGTCTCTGTAACGAA 58380
L L A V H H A V T D G W S N G V L V T

58381 CTGCGCACCGGCTACCGGAACTGCGCGCGGACGCGCGCGCGCGCGCGCGCGCGCG 58440
L A T G Y R E L R A G R P D R R P A P P

58441 GTCCAGTACGGCGACTACGCGCACTGGCAGCGGAGCGGCTGACCGGCGCGCAACTGCGG 58500
V Q Y G D Y A H W Q R E R L T G P E L R

58501 GCCCTGGAGGACTACTGGCGCACCGCGGTACGCGACCTGCCAGGACGAGCTGCCACC 58560
A L E D Y W R T A V R D L P R T D L P T

58561 GACCGCCCCCGCCCGCCGCGCGCGCGAGGGGCCAACACGCCCTGCTGCTCTCG 58620
D R P R P A A R R G E G A N H A L L L S

58621 CCGGAGCTGACCGCGCGCTCGCGACCTGCGCCGCGCGAGGGCGGGTCTGCTGTTTCATG 58680
P E L T G R L A D L R R R E G G S L F M

58681 CTCGTGCTCTCCGCGCTCCTGCTGCTGCTGCGTGGCACCGCGCGCGGACCGGCTCGCC 58740
L V L S A L L V V L R G T G G R D R L A

58741 GTCCGACCCCTCGTCCGCGCGCGCACCGCGCCCGAACTCGAGCGCTCATCGGCTACTTC 58800
V G T L V A G R T R P E L E P L I G Y F

58801 GTCAACGCTCTGCTGCTGCCCTTCGAGACCGGCGCGGACCTCTTCGCGAGTGTGTGG 58860
V N V L L L P F E T G G R T S F A E L W

58861 CGCGGGTCCGCGCGCGGCTGGTGGAGCGTACGCCACCCAGGAAGTCCGCTGGAGAA 58920
R R V R G R L V E A Y A H Q E L P L E K

58921 GCCCTGGAGCTGCTGCGCGCGGACCGCACCGCCCCCGCGACCCGCGGTGCGGCTGCTG 58980
A L E L L R A D G T A P A D P P V G V V

58981 TGCGTCGCCACGAGCCGCGCCCCCGGATCACCTGCCCGGACTCGACCGGAGCTCGAG 59040
C V A Q Q P A P A I T L P G L D A S V

59041 GACGTCGACCTGGGACCGCGCCAGTTCGACCTCGTCTGCTGAGGTGCGCGAACGGCCGAA 59100
D V D L G T A Q F D L V V E V R E R P E

59101 GCGCTGCAGATCGCCTTCCAGTACGACCGGGACCTGTTCGACGCGGCCACGGTCCGGCTC 59160
G V Q I A F Q Y D R D L F D A A T V R L

59161 CTCGCGGACACGTCGACGCGCTCCTCGACAGGCGCGCGCGACCCACCTCGCCCTGT 59220
L A D H V H A V L D Q A A A D P T L P C

59221 GCGGAGCTGCCCGCCCCCGCGCGCCCCCGCGCGCGCGCGCGCGCGCGCGCGCGCG 59280
A E L P A P P A P A A P A R T A G A T T

59281 CTCGACGCCCTGTTCGAGTCCGCGCGCGCGAGAGCGCGCGCGCGCGCGCGCGCGCG 59340
L H A L F E S R A A K S P D A V A L V D

59341 GCGCGCCACCGCGTCACTACCGGACCTCAACACCGCGCGCAACCGGCTCGCCCGCCAC 59400
G H R V T Y R T L N T R A N R L A R F

59401 CTGCGCGCGGTGCGGCTGCGTACCGAGGACCGGGTGGCGTGCCTGCCCCGCGGCACC 59460
L R A V G V R T E D R V A L R L P R G T

59461 GACCGGTGACCGCCACCTCGCGCGCCTCAAGCGCGCGCGCGTACGTACCCCTCGAC 59520
D A V T A T L A A L K A G A A Y V P L D

59521 CCGCGCCTCCCGAGGAACGGCTGACCCCGCTCCTCGCCGACCGCGCGCGCGCGCGGTC 59580
P A L P E E R L T R V L A D A R F A V V

59581 CTCACCCCGCGTATCTGCACGACCGGTCCGCGAGATCACCGCCACCGCGCGCGCGTAC 59640
L T P A Y L H D R S A E I T A H A G H D

59641 CTCAACCTCCCGTCCACCCGACAACTCGCCTACCTCTCCACACCTCGGATGCCAC 59700
L N L P V H P D N L A Y L L H T S G C T

59701 GGCACCCCGAAGGCGCTCCTCGGCAACCAACCGGGCGCGGTCAACCGCGTCACTGGATG 59760
G T P K G V L G T H R G A V N R V D W M

59761 AGCACCGCGTACCGGTTCCGGACCGCGACGTGGCGGTGCCCGCACCGCGCGCGCGCTT 59820
S T A Y P F R T G D V A V A R T A P G F

59821 GTGAGCGCGGTCTGGGAACCTTCGCGCCCTTGCGCGCGCGGTCCCCCTCGTCTCTCTG 59880
V D A V W E L F G P L A A G V P L V L L

59881 CCGACGACGAGGCGCGGACCCGCGCTGCTGACGCGCGCTGGAACGCGCACCGGCTG 59940
P T D E A R D P A L L T A A L E R H R V

59941 AGCCGAGGTGAGCGTCCCGTCTGCTGACCATGCTCTGGACGAGTCCGCGCGCGCG 60000
S R M V T V P S L L T M L L D E S A R F

60001 ACGGACCTCGGACCGCGCTGCGCTGCTCGCACCTGGATCACACGCGCGAGCCCTG 60060
T D L G T R L A C L R T W I T S G E P L

60061	CCGCCCGCGCTCGCCCGCGGGTCCACGACCGCCTGCCCGGCCACCCCTGCTGAACCTG P P A L A R R R F H D R L P G R T L L N L	60120
60121	TACGGCTCCTCCGAGACCGCCCGCGACGCCACCGCGGCCCGCATCGACCCCGCGCCCGG Y G S S E T A A D A T A A R I D P A P G	60180
60181	ACTGCGCTCCCGAGCGGTCCCCGATCGGCACGCCCATCACC CGCTCAGCGCCCTCGTC T A L P E R S P I G T P I T G V S A L V	60240
60241	CGCGCCCGGACCTGCGCCCGCTGCCCGCGCTGATGCCCGCGAGCTGTATCGCGGGGGC R G P D L R P L P A L M P G E L Y A G G	60300
60301	GCGTGCCTGGCCCGCGGTACCAACGCCCGCTCGGCCGAGACCGCGCGCGCTTCCCGCG A C V A R G Y H A R P A E T A A A F P P	60360
60361	GATCCCGACGCGCGGCCCGCGCCCGGATGTTCCGTACCGGTGACAGSGCCCGGCTGCGG D P D G G P G A R M F R T G D R A R L R	60420
60421	GCCGACGCGCGCTGGAACCTCTGGGCGCGGTGACCGCGCAGGTGCAGATCCGCGGCCG A D G R L E L L G R V D R Q V Q I R G Q	60480
60481	CGCGCCGAGCCCGCGGAGGTGGAACACGCCCTGTGCGCCACCCCGCGGTACGGGCCGCC R A E P G E V E H A L L A H P A V R A A	60540
60541	GCGCTACGCGGGAACCCCGACGCCACCGGCTGTGGCGGTACGTGCGGCTCGCTCCCGGC A V T A N P D A T G L W A Y V R L A P G	60600
60601	CCGTTGCGCGCGCGCTCCCGCCAGACCGAGCTGACCGCTTCTGCGCGCGACGCTTCCCT P F A A G S P Q T E L T A F L R L R T P	60660
60661	GCCCACTCGTGCCACCGCGCTCACCGTCTTGGACGAGCTGCCGTTGACGCGCACGGC A H L V P T A V T V L D E L P V T A H G	60720
60721	AAGACCGAACCGCGCGGCTGCCCGCCCGACCCCGCGCGCGCGCGCGCGCGCGCGGACC K T D H A R L P A P D P R A G R P A P T	60780
60781	GCCCGCGCACCCCAACGAGCGTACGCTGCGCGACGCTCTGCGCGGGTGCTCGGCGTG A P R T P T E R T V A D V F A G V L G L	60840
60841	GAGGGCGCGGTGCGCGCGCACGACGACTTCTTCTCTCGCGCGGCACTTCTCTCTCGGC E G P V G A H D D F F L L G G H S L L A	60900
60901	GCCCGCAGTTCGCGCGGAACCTCGCGCCCGCGCGCGCTCGGATCGGGCTGAGCGACGT A R S R G G T P R P P R R P D R A E R R	60960
60961	CTTCGCGGCCCCACCGTTCGCGCAGCGTTCGCCCGCGGACCGACGCGCGCGCGCGCGG L R G G P H R R R S V A A R T D A A R P G	61020
61021	ACCGCCCGGAGCACCCCGTTCTGTCACCGACCCCGCGCGCGCGCGCGCGCGCGCGTCCG T G P E H T P F V T D P G A R H E P F P	61080
61081	CTCACCGACGTCGCGCGGCTACTACGTGGGACGCGAGGGCGGGTTCCGCTCGCGCGG L T D V Q R A Y Y V G R E G G F A L G G	61140
61141	GTCTCCACCCACGCTACCTGAGAGATCGAGGCCCGCGGATCGAGCTCGACGCGTTTACC V S T H A Y L E I E A P R I D V A R P T	61200
61201	GGCGCGTTCGCGGGGTGATCGCCCGGACCCCATGCTGCGCGCGCGTGATCCGTCGCCGAC G A L R G V I A R H P M L R A V I R P G	61260
61261	GGGCTCCAGCAGGTGCTCACCGACGTCGCCCGCTACGACGTGGCGGTGACGACCTGCGC G L Q Q V L T D V P P Y D V A V H D L R	61320
61321	GACCTGGAAGAGCCCGCGCGCGGACGCCGACGCGCGCGCTGCGGAGGAGATGCCCCAC D L D E P A R Q R R R A A L R E E M S H	61380
61381	CAGGTGCTGCGCGCGACCTCTGGCCCTGTTCGACGTCGCGCTCTCCCTCGGCCCCAGC Q V V P A D L W P L F D V R V S L G P T	61440
61441	GACGCCCTCGTCCACGTGGGGTGACGCGCTGATCTGACGCGCCACAGCTTCGCGCTC D A L V H V G V D A L I C D A H S F G L	61500
61501	GTCTGCGCGAACCTCGCGGCCGTTACGCGGACCCCGCACCGCGCTTCCGCGCCCTGACG	61560

V L A E L A A R Y A D P A R R F P P L T

61561 GCGGACTTCGGGACCACTCTCCATCAGGAGCGCTCCGGGAACCGAGATACGG A D F R D H V L H Q E A L R G T A Y A 61620

61621 GCGGCGGAGCGGTACTGGCGGAACGCTGCCCGAGCTGCGCGCCGCGCCCGAACTGCC A A E R Y W R E R L P E L P P G P E L P 61680

61681 CTGGCCGTGCGCGCCGAGACCTCGGCAACCGCGCTTACCCGCGCTCGGCGCGGTG L A V A P E T L G T P R F T R R S G R L T 61740

61741 GACCGCGCTCTCGGACGGCGGTCAAGGACCGGCGCGCGCGCGGTCTCAGCCCTCC D A A S W T A V K D R A R R A G L S P S 61800

61801 GGCCTACTGCTGGCGCGTTCGCGGAGGTGATCACCAGGTGGAGCGCGCGCGCTAC G V L L A A P A E V I T A W S G R P R Y 61860

61861 TCGCTGATGCTGACGGTCTTCGACCGCCCGCGCTCCACCCGAGCTCGGCGGATCGTC S L M L T V F D R P P L H P D L G R I V 61920

61921 GCGGACTTCACTCGCTCAGCCTGCTGGAGGTGACCAAGTCCGCGCGCGAGCTTCACT G D F T S L S L L E V D H S R P G D F T 61980

61981 GACAGGCGCGCGCCCTCAGCGCGCTTGGCAGGACTCGACCACTGGCGCGTGGC D R A R A L Q R R L W Q D L D H L A V G 62040

62041 GGCCTGACGGTGACCGGGAACGGGCGCTGCGCCAGACGCCGAGCCCGGTCTGCTCA C G V T V T R E R A L R H D A R P G L L A 62100

62101 CCCGTGCTTTCACCTCCGACCTGCCTGTGCGGAGAGCCGCGCGGAGGAGCGGACGGG P V V F T S D L P V G E T A A E D A D G 62160

62161 GGAGAGGATGGCGCTCGGAGAGCCGCTTACGGCGTCAAGCAGACCCCGAGGTCCAT G E G W A L G E P V Y G V S 62220

62221 CTCGACCATCAAGTCGCGGAAGACCGAGGGAGTTGGTCTCAACTGGAGCGCCGCGAA L D H Q V A E D R G E L V F N W D A V E 62280

62281 GACTGTGTCGCGCGCGCGCTTGGACGCACTGTTGCGCGCTACACGCTCGCTGAC C D L F A P G A L D A M F A A Y T A S L T 62340

62341 CGCTGGCCCGGAGCCCGCAAGCCTGCGCGCGCGCGCGCACTCCGCGCTGCCACCG C R L A R S P E A W R R P G T P P L P T 62400

62401 CAGGCGCGCGTGGCGCGCGCACCGCGCGAGCGAGCGCGCTGCCCGCGCGCTGTG Q A A V R R R T A A T E A P L P A R L L 62460

62461 CACGAGCGCTCGGCGAGCGCGCGCGCGCACCGCGCTGACCGCGCTGTGTCAGCGG H E A V G D A A R R H A D L T A L V D G 62520

62521 GACACCGGATGACCTACCGCGCACTGACCGAGCAGCGCGCGCGCTGCGCGCGAGCTG D T R M T Y R R L T E H A R R V G R T L 62580

62581 CGCGCTCTCGCGCGCGCGCGCGCTTGGTCCGCTGGTCCGCGCAAGGGTGGCGG R R L G A R P G R L V P V V A R K G W R 62640

62641 CAGGCGCTCGCGCGCTGGCGCTCTGGAGTCGGGCGCGCGCTACTGCCCTGGACCCC Q A V A A L G V L E S G A A Y L P L D P 62700

62701 GAACTGCCCGCGGAACGGCTGCTCCACCTGTAAGCGCGCGCGAAGCGCGCTCTCTCT E L P A E R L V H L V R R A E A A L L L 62760

62761 ACCGAACGCGCGCTGCTGGACAGCTGCGCGTCCCGCTCGCGCTCACCCTGCTCGGTG T E R A L L D T L A V P V G V T V L A V 62820

62821 GACGACGAGCGCGCTCGACGCGAGCGCGCGCGCTGAGAGCGTGCAGAACCTCACC D D D A A L D A D G G P L Q S V Q N L T 62880

62881 GACCTGGCGTACACCATCTTCACTCGGGCTCCACCGGCGAACCAAGGCGCTCATGATC D L A Y T I F T S G S T G E P K G V M I 62940

62941 GACCACCTCGGCGCGCGCAACACCTTGAATGCGTCAACCGCGCTTCCGACCGGCCCC D H L G A A N T L E C V N R R F G T G P 63000

63001	GGCGACGCGGTCTCGCCGCTCTCTCCCGAGCTTCGACCTCGCGTCTACGACCTGTTC G D A V L A V S S P S F D L A V Y D L F	63060
63061	GGCGTGTGGCCGCGGCGGCACCGTGGTCTGTCGCCGCCACGACCGCGCGGACGCC G V L A A G G T V V V P A H D R R R R D P	63120
63121	GGACACTGGGCGGAGCTGATCCGGCGGAGCGGGTCACCTCTGTGGAATCGCTCCCGCG G H W A E L I R R E R V T L W N S V D F A	63180
63181	CTGGGACACCTGCTACCGAGTACGCCGAGGCCCTCGCCCGGACGCGCTTGGCACCTG L G T L L L T E Y A E A L A F D A L R T L	63240
63241	CGGCGGGTCTCTCAGCGGCGACTGGATCCcctcggaactgccgacggatCCCGGCC R A V L L L S G D W I P L G L P D R I R A	63300
63301	CTGTCCGCCCCCGGCGGCCACCGTGATGAGCTCTGGGCGCGCGACGGAAGCTCCATCTGG L S A P G A T V M S L G G A T E A S I W	63360
63361	TCGGTCTGGTACGAGATCGGAAGGTGCACGAGGCGTGGAGCAGCATCCCTACGCGACC S V W Y E I G K V H E A W S S I P Y G T	63420
63421	CCCATTGGCCACACGCGGCTGGAGGTCTCTGACGAGCAGCTCGCGGCCCGGCCGACTGG P M A N Q R L E V L D E Q L R P R P D W	63480
63481	GTGCCCCGCGAGCTGTACATCGGCGGCGACCGGCGTCCGCAAGGCTACTGGCGCGGACCG V P G E L Y I G G T G V A K G Y W R D P	63540
63541	GAACAGACCTCCCTGCGCTTCCCGGTCCACCGGCGAGCGGGCAAGCCTGTACCGCACCC E Q T S L R F P V H P G S G Q R L Y R T	63600
63601	GGGACTCTCGCCCGCACCTCCCGGACGGCACGCTGGAATTCCTGGGCGCGGACGACGAC G D F A R H L P D G T L E F L G R Q D G	63660
63661	CAGGTGAAGATCGGCGGATTCGCGGTGGAATCTGGGCGAGGTGAGGCGCGCCTCGCGCGA Q V K I G G F R V E L G E V E A A L G R	63720
63721	CTGCCCCGCTGCGCCCGCGCGGTGATCGCCACCGGTGACCGCGCGGCGGCGACCGCGC L P D V A A G A V I A T G D F L G R Q D G	63780
63781	CTCGTCTGGTCTCGCCGTACCGCCCGCGGAGGCGGCTTCGACGCGCGCGGCTCGGACGG L V G F A V P A R E G G F D A A G L R R	63840
63841	CAACTGCCCCGCGGTGCGCGCTTACATGGTCCCCAAGCCTGTGCGCCTGGACCGG Q L A R R L P A Y M V P T T L L P L D R	63900
63901	CTGCGCTGACCGCCAAAGTCAAGTCAAGCGCGCGCACTCCAAAGCCTCGTCCCGCGC L P L T A N G K V D R A A L Q R L V P G	63960
63961	CGCGCACCGGCGCGCGGGAACCGGCCACCGCCCACTCGCCGTTCCCGCGCGCTGCCCC R A P A P A E P A T A P P A R S R A V P	64020
64021	GTGCCCCGCTGGCTCGCGGACCTGTGGTCCGAACCTCTGACGTGCGGAGCGCGACCCC V P G W L A D L W C E L L D V P E A D P	64080
64081	GACGCGAATCTTCTGCGCCTCGGCGGCACTCCCGGCTCGCGATACCCCTGGTCAACCGG D A N F F A L G G T S R V A I T L V R T	64140
64141	ATCGAGGCCGACTCGCGTCCGCGTCCCGCTCGCCCGCTCTTCGACGCGCGCACCTG I E A R L A V R V P L A R L F D A R T L	64200
64201	GGCGGCTCGCGGAGACGATCGCGAATGTTCGGCGCGCGCGGAGGAGGACCGCGCACCC G G L A E T I A E L S A A E E P A P	64260
64261	GCGGAGCGCGGTGTAAGCCCGGACCGCGCCACCGCGGAGCGGCTTCGCGCTCACCGAC A E P V Y A P D P A T R H E P F P L T D	64320
64321	ATCCAGCGCGCTACTGGCTCGGCGCGGACCGCTCCCTCTCCCTTGGCGGCGCGCACG I Q R A Y W L G R H R S L S L G G V A T	64380
64381	CACACCTACTCGAATCGAGCTGAGGACCTCGACCCCGCGGCTCCGACGCGGCTC H T Y L E L D V E D L D P G R L Q T A	64440
64441	CGCCGCGTATCGACCGCCACGACGCGCTCCGCGCTCGTGGTCTCTCCCGGACGCGCGGCA R R L I D R H D A L R L V V L P D G R Q	64500

64501 CAGATCCTCGGCGACGTACCGCCGTACTCTCTCGCCCCACCGACCTGCGGGGCGAGGGG 64560
Q I L G D V P P Y L L A H T D L R G R A

64561 GAGCGCGAGGCGGAACCTGGCCCGCTCCGCGAGCACATGTGCGACGAGGTGCGCGACGCC 64620
D A E A E L A R V R E H M S H E V R D A

64621 TCCCGCTGCGCGCTGTTGACGTACGACCCACCGCTGCGACGCTCGCGACCCCGGCTG 64680
S R W P L F D V R T H R L D D V R T R L

64681 CACCTGAGCTTGGACCTGCTCATCGCGACGCCACAGCGTCCACGTACTCACCGCGGAC 64740
H L S L D L L I A D A H S V H V L T G D

64741 CTGCTCACCTTCTACGCGACCCCGACGCGGCCCTCGCGCCCTCGGCTGCTCTTCCGC 64800
L L T F Y A D P D A A L P P L G C S F R

64801 GACTACGTCTGGCCCTCCGCGCCACGCGGAGGCGAGCCGCGCGCGCGCCCTCGAC 64860
D Y V L A V R A H A E G E P R R R A L D

64861 CACTGCGGGCCCGGCTGGCCGACCTGCGGGCCCGCCGCTCGCGCTCGGCTGCGG 64920
H W R A R L A D L P G P F G L P L R C R

64921 CCCGAGGAGCTGACCGCGCCCGGTTGCGCGCTCACCACCGGACTCGGCCCCGACGCC 64980
P E E L T A P R P A R L T T G L G P D A

64981 TGGGACAGCTGCGGCGCGCGCGCGCGCGGCGGCGGCGGCGGCGGCGGCGGCGGCGG 65040
W A R L R R A A A A A E L T P A A L I C

65041 GCGCGCTTCTGCGACGCTCTCGCCAGTGGAGCGACACCCCGCTTCACTTCAACCTC 65100
A A F C D V L A Q W S D T P R F T L C N L

65101 ACCACCTTCCACCGCCCGCCCTGCTCCCGCGGTGGACGACCTCGTCGGCGCATTCACC 65160
T T F H R P A L L P G V D D L V G D F T

65161 ACCACGACCTGCTCGGGTGGACGCGGAGGGGACACCTTCCGGGACCGGCGCCCGCGA 65220
T T L L G V D G E G D T R A R R

65221 CTCAGGACCGCATCTGGGAGGACCTCGAACACCGGTCGTACAGCGGTCGAGGGTCTG 65280
L Q D R I W E D L E H R V V S G V E V L

65281 CGGATGCTGCGCGGAGCGGGGACCCACGACGCGCTCGGATGCGGTCGCTCTTCAAC 65340
R M L R R E R G T H D A V R M P V V F T

65341 AGCACCTGCGGGCCGCGGCCCGCCCGCGGACGCGCCCGCGCTCGCGGGTACGG 65400
S T L R A A G P A P R T A P P A W R V R

65401 CCGCGCTACGCGATCAGCCAGACCCCGCAGGTCTCTGCTCGACCATCAGGTGAGCGAGC 65460
P G Y A I S Q T P Q V L L D H Q V S E S

65461 GACGCGATGCTGCTSCACTGGGACTACGTGCGGAGCGCTACCGCGCGGCTGATC 65520
D G R L L V C T W D Y V A D A Y P P G L I

65521 GAGGCCATGTTGGGGCCTTCCGAGCGCTCTCGCGCTCGCTCGCGGTCACGACGACG 65580
E A M F G A F E A L L A S L A G H D D D

65581 GCGGCCACGACGACGACGCGGCCACGACGACGCGCCCGGCCACGACGCGGCCCGGC 65640
A G H D D D A G H D D G P G H D D G P G

65641 CACGACGACGCGCCCGGCCACGACGACGCGCCCGGCCACGACGCGGCCCGCGCGAC 65700
H D D G P G H D D G P G H D D G P G R

65701 GACAGTGCAGATCACGGCCACAGTGCCACGACGACGACGCGCCCGCCAAACGACAGA 65760
D S A D H G H S A T H D D S A A R N D R

65761 GAGGAGGTGACGCGAGTACGAGCGCCCGGCCACGCGCACTGCTCCCCGCCGACC 65820
E G G G P E *
M T S A R P T P T L L P A D Q (orf16)

65821 AGCGGAGCTGCTGCGGATGATGAACGACCGCACCGCACCCGTCGCCGCGCACCCCTCA 65880
R E L L R M M N D R T A P V P A H T L T

65881 CCGCCCACTGGCGGACGCGCGCGCGCACGACGCGCGGCTCTGGCACTGGTGGACCGG 65940
A Q L A D A A R T H D R A L A L V A P G

65941 GTCTGACACTGAGCCACGCCGAACCTGGACGCCCGGGCGGCGTGGCGCGCCCGCTCA 66000
L T L S H A E L D A R A A A V A A R L T

66001 CGCGCGCGGGCGTCATCCCGGGGACCGGTCGCGCTCGCGTCGAGTACGGCTGGGAGC 66060
A A G V I P G D R V A L A V E Y G W E Q

66061 AGGTCGTGGGCGCCCTGGCGCGCTCCGCGCGGAGCGCTCGCTGCCCTCGCCCCCG 66120
V G G A L A A L R A G A V C L P V A P G

66121 GGCTGCCCGCGCCCGCTGGCAGCAGCCACCGGGCGGGGCGAGCGCGCTCCTCA 66180
L P R P A R W Q H A T R A G A T A V L T

66181 CCCAGTCTTGCTCACCAGCGCATCGACTGGCGCAGGAACCTGCCGCTCTCTCCGTTG 66240
Q S W L T Q R I D W P Q E L P V L S V D

66241 ACGAACCCGGGCGCGCGTACCACCCACCGCCCGGCGCAGCAGCGGTCCGCGACCG 66300
E P G P P V P P T T A P A D G R S A T D

66301 ACGCGCGCTACCGGCTGGAGCCCCCGTCAGCCACGCGCGATCACCAACCGCGCCCTGG 66360
A A Y R L D A P V S H R A I T T A A L E

66361 AGATCGACCGCGCTTCGCGCTCGGACCGGCGACCGGCTCTGCCTGGCCCCCGCG 66420
I D R A F R V G P G D R L L A L A P A D

66421 ACTCGCGCTGCTCTCTACGAACCTGTTGGGCGCCCTCCTGGCGGATGCGCGCCCTGTC 66480
S P L A L Y E L F G P L L A G A A L V L

66481 TCACCGGGGACATCGACCTGCGGATCCCGGAGCCCTGCACGAGCGCTGCGCACCCACG 66540
T R D I D L R D P G A L H E A L R T H G

66541 GCGTCAACCTCTGGCACTCGCGCGCCCGCTCTCGGCTCTCTCGACCACTCGCG 66600
V T L W H S P P A L L G L L D L H A D

66601 ACCGGGCGGCAAACTGCGCGAGTGGCTCCGCTGGTGTGCTCGGCGCGGAACCGCTCG 66660
R G G K L P E S L R L V L L G G E R L D

66661 ACCCGCGCTGCTCGCGCGGTCGCGAGAGCGCCCGCACGAGCGCGCGCTGACCCACC 66720
P A L V R R V R E S A P H Q P A V A H C

66721 TCTCTCGGCCACCCCGTCGCGCGCCCTGGACCACTGCTCGGAGACCGCGACCTCGCC 66780
S A T P S G P W T T C L E T G D L A P

66781 CGGAATGGCGCTCGTCCCGTCGCGCGCCCTGCGCAACCGAGCGGGCGACATCTGT 66840
E W R S V P V G A P L P N Q R A H I L S

66841 CGGAGACCTGCGCGCCCTGCCCGTCTGGGTACACCGGCGGCTCCACTACGGCGCGTGC 66900
E T L R P C P V W V T G R L H Y G G V A

66901 CGCGGAGCGCCCGACCGGAGAGGAGCAGCACCGCGACCGTCCCGCACCCCGGAGACCG 66960
A E P P T G E E H A P A T V P H P E T G

66961 GCGAACCGCTGCTGCGCACCGGCTGTTGCGCGCGCTGCTGCGCGAGGCGCTGATCGACG 67020
E P L L R T G L F A R L L P E G L I D V

67021 TCGTGGCGAAGAGACCGCGCGGATCAGCGTCCGCGACCGGCGCCCTGAACCTCCAGGAC 67080
V G D E T A R I S V R D R P L N L Q D T

67081 CGGAGACCGCCTCGCGCGCCACGAGGACGTGCACTCGCGCGTGGTGGTCCCGCTCGGGC 67140
E T A L A A H E D V H S A V V V P V G R

67141 GGGGAGACGAGTGGCTCGCGCGGATACGGCTCCACCCCGCGCCACGGCGCGCCCGACG 67200
G D E S L A R V R L H P G A T A G P D E

67201 AACTCTCGGCCATCTGCGCGCGAAGTCTCCCTTACCTGCTGCGCGCGCACATCGAGG 67260
L L A H L R R K V S P Y L L P G H I E V

67261 TGGGCGCTCGCTGCGCTCACCCGGACGGGCGGTCGAGCGCGCGCTACCGCGG 67320
G G P L P L T R D G R V D R A R V T A E

67321 AGGCCCCCGCCCGCTGCCGTGCCCGCGCGCGCGCGCGTGGGACCGCGCGCGCG 67380
A P A P A A V P A A A P A A S A P A R D

67381 ACGAGGCGGAACCTCTCGCCCAAGTGGCGCGGTCACCTGCGGCTGCTGGGAATCGGCG 67440
E A E L L A Q V A R V T C R V L G I G A

67441 CCGTGAACCCGATATGAACCTGCTGACGCCGGTGCCACCTCGTGAACCTCGTCCGCC 67500
V E P D M N L L D A G A T S V E L V R L

67501 TGCGACCCGCTCTGGAGGAGAACTCGGCCTCGACACCGACATCGAGGAACCTGTCGCCT 67560
A T A L E E E L G L D T D I E E L L A F

67561 TCCGTCGGTCCGCGTATCGTCCGCCCGACCTCGGCCGCCGACGACACCCAGGCC 67620
P S V A V I V G R H L G R R T A P P A R

67621 GGGACCCCGTCCGCCCGCGTCCGTAGCGTTCGACCCGGGTCGTAATCGCCGCCGCC 67680
D P L P P A S V A F A P G S V L P A P P

67681 CCGCGCCCGGACCGGTGCCGCCCGCGTCCGTGCCGCCCGACCCCGGTCCGTACCGCCCG 67740
A P G P V P P P A S V P P A P A S V P P A

67741 CGTCCGAGTCCCTACCGCTCGCGCGCCCGACCCGGGCCGTGCCACCCACGCCCTCC 67800
S E S S P L A P P A P G P V P P T P V V P

67801 CGCCCGCCTCCGTCGCCGCCCGTCCGGGCCGCCGCCGACGTACCGCCGCCGCCGCC 67860
P A S V P P A S G A A P H V P P A P P A

67861 CACCATCCCGCGCCCTCCGTGCCCGGCCGCCGCCGCCGCCGCCGCCGCCGCCGCCGCC 67920
P I P A P S V P P A P R P Q P P L L T G

67921 gcatcgccgcccgcagcgTTCAAGGACGCCACACCGCATCCGGCAGCTTCGACG 67980
I G A R Q A F K D A H H G I R H E F D A

67981 CCACGACCGGTCCGCCCTCAGCGGCCCGGACGACACCCCTCAGCCGCCGTCCGACG 68040
T D G V A L S G P D D H H L T A R R S H

68041 ACCACCGCTTCGACCCCGGCCCGTGAAGCTGCCGGACCTGGCCGCCCTCTCGGCCGCC 68100
H R F D P G P V T L P D L A A L L G A L

68101 TCCGCCCGGTCCGCCGCCCGGAGGCGAACCCAAATACGCTATCCGTCCGCCCGGTTCC 68160
R R V R G P G G E P K Y A Y P S A G S S

68161 CTTACCCCTCCAGACCTACCTGCTCGTCCACCGGGAGGTCGACCTGCGACGCGCG 68220
Y P F Q T Y L L V H P G K V T G L P G G

68221 GCAGCCACTACGTCCACCCCGCGCGCAACCGCTGGTGAGCATCGACCCCGCGACCC 68280
S H Y V H P A R N R L V S I D P T A T L

68281 TGCCCGCCGACGCGCACGCGGAGATCAACCGCGCGCCTACGGGGAGCGGCCCTCTCC 68340
P A D A H A E I N R A A Y G E A F A S L

68341 TCTACCTCATCGCCCGATCGACGCGATCACAACCGCTCTACGGCGATCTCTCTCGGACT 68400
Y L I A A I D A I T P L Y G D L S W D F

68401 TCACCGCTTCGAGGCGCGTGCCATGACCCAGTTGCTGATCGGACCGCGCTCGGACCG 68460
T V F E A G A M T Q L L M R T A V G T G

68461 GCATCGGCCTGTGCCCGCTCGGCACGATGGACCCCGCGCGCTGCGCGCGCGGTTGCGCC 68520
I G L C P V G T M D P A P L R R A F A L

68521 TCACGACCGGCAACCGCTTCGTCCACGCCCTCTCGCGGGCGCGGCCCGCACGAGGCC 68580
T D R H R F V H A L L G G R P R T E A P

68581 CGTGAACCGGACCGGCCCTGGCGGGCGCGCGCAGAGCGTGCACACCCCGACCGCCCG 68640
M N R R H G P L A G R R Q S V D T C A S A C A
(orf15)

68641 GTGGGTGGCGGACGGGCAACCGGGGCTGCCGCTGGAGGTGGCCGCCACCCGGGACGG 68700
W V A P T G T P G L P L E V A A T R D G

68701 CGTCCACCGGCGAATGGGCCCGCACCCCTCGACACCGTACCGCGCTGGCTGACCGG 68760
V D P A E W A R T H L D T V T G W L H R

68761 TCACGAGCCGTCTGTTCGCGGCTTCGGCGTCCGCCCTCGACCGCTCGGCGACGTGT 68820
H G A V L F R G F G V G L D G F G D V V

68821 CCACGCCCTGGCGGATCCCGGAGGCGTACGTGGAACGGTCTGCGCGCGCACCCGCCCT 68880
H A L A G S P E A Y V E R S S P R T A L

68881 CGGGCATCACCTCTACACGCCACCGACCACCCCGCGACACGCCCATCCCCCGCACAA 68940
G H H L Y T A T D H P A D Q P I P P H N

68941 CGAGAACTCTACCAACTCCGCTTCCCGGACGGCTGGTCTTCGGTGCCTCACCCGCGC 69000
E N S Y Q L R F P G R L V F G C L T P A

69001 CGGACCGCGCGCGACCCCGCTCGCGACACCCGCGCGCTCCTGGGCGCGCTGCAGCC 69060
R T G G A T F L A D T R R V L G R L D P

69061 CGCCCTCGTGGCGCCTTGGCCCGCGCGGGTGTCTTACAGCGCAACTACGCGGACGG 69120
A L V A A F A R R G V L Y Q R N Y G D G

69121 GATCGGCATGTCTTGGCAGGACGCTTCCAGACCCGCGACAGGCGGCTCACCGCCTA 69180
I G M S W Q D A F Q T R D K A A V T A Y

69181 CTGCGCGCGCCCGCGCTGACGTCGAATGGAAACCGACGCGGGGTGCGGACCAACCCA 69240
C A A R R V D V E W K P D G G L R T T Q

69241 GGTCCGCGCCCGCTCGCGTCCACCGGCGACGGGGAGCGGGTGTGGTTCAACCAAGC 69300
V R P A L A V H P A T G E R V W F N H A

69301 CGCGTCTTCCAGCTCTCCGCCGCGCGCGCGCTGCGGACCGCCTGCTGGCCAGTT 69360
A F F H V S A R P F A L R D A L L A Q F

69361 CGACGAACGCGACCTGCCGAGCCACTCTGTCTACGCGACGGCGCGCCATCGAACCCGC 69420
D E R D L P S H S C Y G D G R P I E P A

69421 CGTCATGGAGAACTGACCAACGCTACGCCCGCAACTGGTGGCGCGCGCTGCGCGGC 69480
V M E E L H H A Y A A E L V A P A W R A

69481 CGGCGACGCTCTCCTCGTGACAACTCTACCGCGCACGGCAGGAAACCTTACCGCG 69540
G D V L L V D N L L T A H G R E P T T G

69541 CGAACGCGCGTCTGTCGTGGCATGGCACAGCGCTGGAATGGGACGAGGTGAGCGCTG 69600
E R R V V V G M A Q P L D W D E V S A *
M (orf14)

69601 ACCGCCCCCGGCACACCGCTGCCCGGACCTTCTGTCAGCGCGCGCTGTGGCCGCTCACT 69660
T A P G T P L P A T F V Q R G L W F S T

69661 CGCCACCCCGCCCGCGGAGGTACCCACGTCGCGGCCCTGCGCTGACCCGGACACCC 69720
R H A R P A E V T H V R A L R L T G D T

69721 GACACGCGCGGCTCACCGAGGCGCTCGGCGGGTACCGCCGCCCTCCCGCCCTCACC 69780
D T A R L T E A V R R V T A A L P A L T

69781 GCCGAATCTCGGCGACGAGAAACCCGCTGACCTCGCGCGGACGCGCCCGAGGTC 69840
A E L S G D E E P R L T L R P D A P E V

69841 ACCCGGTGACCTGCGCGGAGCCCGTCCCGCGGACGCGCGCTGCTGCGTGGCGCTG 69900
T P V D L R G A P S A G R D A V C V A L

69901 CTGCGCGCGACCGGACCACTCTCGCGCGGACGCGACCGGCGCGCTTCCACCTGGTG 69960
L R A D R D H P R A G R H R A R P H L V

69961 CGGCTCCACGACGAGAGCGGTGCTCGGCTACGCGGCCACACCTCTCTCTGACACA 70020
R L H D D E T V L A L T A H T L L L D T

70021 CGGTCTCTTACGCGGTGCTCGCGCGCGTCTGCCAGGGGTACGCGCGCGCTTCCGCCCC 70080
P S L Y A V L G A V C Q A Y A G R F P

70081 GAGCACTACCGGACGCCACACCTGCCCCGACGCGCCCCACGCCCCCTCTCTCGGTGG 70140
E H Y R D A T T L P D A P H A P L S G R

70141 GCCCGGCGCTCCGCGCGCGCTGGTGGACCGCGCGCTGGCGCGCTGCGCGCGCGGCG 70200
A R A S R R R W W H R R L A A L P G P A

70201 CGGCGCCCCCGCGCCCGCCCCGCGACCGGGTGACCGAAACCCACGCGTGGCGATCCCC 70260
P A P A G P P R D R V T E T H R L R I P

70261 GCAGCGCGTGGAAAGCCCTGACCGCCCTGACCGCCCTGGGCGGCCCCCTCGGCGGCAAC 70320
A A R W K A L T A L T A L G G P L G G N

70321 GGCTCGCTGCGCGTATGAGCCCTGGCGCGCTGGTGCCTGCGCGCCCCGGAACACCGGG 70380

G S L A V M A L A A W C L R A P D H R G

70381 CCGGCCCGCTTACCACCGTCGTCGACCTGCGCGACCACTCGGACTCGGGCCCGCGCTC 70440
P A R F T T V V D L R D H L G L G P A T C

70441 GGCCTGCTACCGACCGCTCTGCTTGGGCGCGACCTCGGCGAAGCGCGCGCCCTCC 70500
G P F T D R L V F G A D L G E A P R P S

70501 TTCCGGGAGCTACGCTGCGCGCCCACTCGGCTTCTGGAGCGCGCTCGTACCTACCTC 70560
F R D V T L R A Q S G F L D A V V H Y L C

70561 CCGTACGGCGAGTCGCTGGAACTCGGAGGAACTGGGCGCGCTCACCGCGCCCGGACCC 70620
P G D V V E L G R E L G R V T A P R T

70621 GCGCGCACTGGGACGTGGGCTGAACCTCTGCGCAACCGCGCCACGAGCGCGCCACC 70680
A A H W D V A L N F C R N P P T S A A T

70681 CCGCGGCAACGCACTCTCGCGAAGCGCGCTGCTCACTGAGCTGTTCCGCGAGGCGGAC 70740
R G E R T L A E R G L S I E L F R E A D

70741 CTGCTCGGCGCGCGCGGACCGGTCTCCGCGCACCGGTGGGACGCGACGCTGCTCGCCCTC 70800
L L G A A G T G P A H R W D G T V L A L

70801 TCCTAGGCGAACTCGGCGAGGACACCGTCTGCTCTCGACCGGACCGCGGACCGCG 70860
S L G E L G D D T V L V L D A D R D H P

70861 CACCACGGAACCGCGGACCGGCTGCTCCACCGGATGGAGGAGCGCTCTGGCGGCGCTC 70920
H H G T A D R L L H R M D E A L L A Q V

70921 GCGGACCCGAGCGCCCGCTGCGCCCTTGGCGCGCGCGCACACCGAGGAGCCAC 70980
A D P D A P L P P L P A P A H T T R S H

70981 CGATACCAACGACCCCGGAGCGCGCGGAGCCACCTACCACTGGTGGTCAACGACG 71040
M T T P R T A A E P T Y H V V N D E (orf13)
R *

71041 AGGAGCAGTACTCGATCTGGCTCGCGAAGCAGGAGATCCCGGCGGCTGGCGGCGCACCG 71100
E Q Y S I W L A E Q E I P A G W R A T G

71101 GAACCTCGGCGACCCAGGAGGAGTGCCTGCGCACATCGAGGAGTGGAGCCGACATGC 71160
T S G T Q E E C L R H I D E V W T D M R

71161 GCGCCCGGAGCTGCGCGGAGGCGCATGGCGCGGCGGAGCACGCGGAGCCCGCTCCGCGCC 71220
P R S L R E A M A A A E H A E P A P A P

71221 CGGCGCGGCGGAGGAGGCGGAGCTGCTGAGCGGCTGCTGCGGCGGCGACGACCGCG 71280
P A A E E E P S L V D R L C A G D Q P V

71281 TGGAGTGGCTCTCGCGCGGAGCGCACCGGCGCGCGCTGCGGAGGCGCTGACCGCG 71340
E S V L R P E R T A A A L R E A V D R G

71341 GCTAGCTCTTCTGCTCGGCTTGGCGGCGACCGCGGCGGCGGCGGAGCTCGGCGTGG 71400
Y V F V R F A A T R G G T E L G V A V D

71401 ACCCGCGCGGACCACTGAGCGGCGACGAGCTGCGGCTGAGCGGCGACCTCTACCTCT 71460
P A A T T M D G T E L R L T G T L D L

71461 ACTTGAACCGGTCGCGTGCACGCGCGGCTGACCTGACCACTTCAAGGCGGAGGCG 71520
F E P V R C H A R V D V T T F T G E G R

71521 GCCTGGAGCGGCTGCTCGGCGACCTGACCGCGCGCGGCGGCGGCTGAGGCGCGGCTC 71580
L E R V S G T *

71581 GGGACCGGCGCGGACCCCTCGAAGGAGGAGACCCCATGACCAACCCCATGACCAACCC 71640
M T T P M T T P (orf12)

71641 CACGACCAACCGCACCAACCCCGCACCGCGCTCTTGGCGCACCTCGCGCGCGCGGCT 71700
T T T R T T T R T A V F A H L R A P G L

71701 CGGCGGCTCTCTCAGCGCAACATCGGCTCTGCGCTGCTCGGCGCGCGCGCGGCGGAC 71760
G D L L Q R N I G L A L V R R A R P A C

71761 GGGCGTCAACCTGGTGTGCGGCGGAGACCTGGCGGCGGCTTGGTTCGCGCACTACCCG 71820
A V T L V V G E D L A A R F G P A L T R

71821 CCACACGTAAGCCACGACGTGCTGCCCTGCCCCAGCGGGGCGAGGCGACCCCGGTG 71880
H T Y A T D V L P C P Q R G E A D P R W

71881 GCCCGCTTCTCTGGCGCACCTTGGCGGACCGCGCTTGGCGCTCGCGCTGTCGACCGGA 71940
P A F L R T L A D R R F A L A V V D P D

71941 CAGCGAGGCGCTGCACGCGGCGCACGCGGGCGCGCGGCTGCCGAGCGGATCGGCCT 72000
S Q G L H A G H A R A A G V P E R I G L

72001 GCGCGAGGACCGGCGCGGACGAAACACATCCCATCCGCTCCGCTCCGACGTCCTCC 72060
P Q D R P G D E H I T H P I R L P R P L

72061 GTGGGGGACCCCGACCTGTACGAGTACGCCACTGCCCTGCCCGCGCTGGGCTGCC 72120
W G T P D L Y E Y A T A L A A A L G L P

72121 CGCACCGCGCGCGCGGCGACGTCTGCGCGAGCTGCCCGCACCGCGGGCTCCGCC 72180
A P P R P G D V L P E L P R T R G V R P

72181 GCGCGGCGGCTGCGCGCTGCCGCTGCCGCTGCCGCTGCCGCTGCCGCTGCCGCTGCC 72240
P T A G L P R P L V A V H P G G A P H W

72241 GAACAGGAGATGGCGCTCAGCACTACGCCCGGCTCTGCCCGCGCTGCCGCGGA 72300
N R R W P L E H Y A R L C A R L A A E L

72301 CTGCGCTCTCTCTGCGCTGGGCGAGAACCGGACCGCGCGGCTGGAACTGCTCCG 72360
S A S L C L L G D E A E R P E L E L L R

72361 GCACGCGCTCTGACGCGGTCGCCCGGAGCGCTGCTCCACTCGAGCGGGCGGCGAC 72420
H A V L T R S P R A V V H L E A G A D C T

72421 CGACCGGACCGGACGCTCTGCGGACGCGGACCTGCTGCTCGGCAAGACTCCTGCT 72480
D R T A N V L A D A D L L V G N D S S L

72481 CGCCGCGCTGCCCGCGGCTGCCGACCGCTGCCGCTGCTCTACGCGCTCGACCGG 72540
A H V A A A V R T P S V V L Y G T G C A T

72541 CAGTACTCTGTGACGAGGATCTACCGGTACACCGCGGGGTCTCCCTGCGGTGCGCG 72600
E Y L W T T R I Y P Y H R G V S L R W P C

72601 CCAGCGGCTGCCCGGACCGCGGACGCTGCCCGCGGCGGCTGCCGCGGCTGCCG 72660
Q R L R H A A G E L A G R R C A H G C V

72661 CCGTCCCTACCGGGCGCGCGGCGCGGCTGCCGCTGCTGCGGCGGCTGCCGCTGGA 72720
L P Y Q G P A G P Y P R C L A D L P V D

72721 CAGGCTGTCGCGGCGGTGACCGCGGATGGCGGAGCGCGCGCGCGCGCGGCTG 72780
R V W P A V T A R W A S P H P V T I R S

72781 TACCCCATGAGCGCGACCGCTCCCGGCTGCCGAGCATCTCTCCGCTCAACTTCAACC 72840
T P *
M S A D P S R V R T I L S V N F N H (orf11)

72841 GACGCTCCCGCGTGTCTGTGCGGGAGGCGAGGATCGCGGCTACGTCACACCGAGCG 72900
D G S G V L L R E G R I A G Y V T T E R

72901 CGCTCCGCGCTCAAGAAGCACCCGCGCTGCCGCGAGGAGGACCTCGAGAACTGCTG 72960
R S R L K K H P G G L R E E D L A D L L A F

72961 CAGGCGGGGCGGACCTCTCGACATCGACCACTCATGCTCTGCAACTGCAACCATG 73020
Q A G A D L S D I D H V M L C N L H T M

73021 GACACCGGACATACCGGCTGCACGCTCCGACCTCAAGGAGACCTGCTCGGCTTC 73080
D T P D I P R L H G S D L K E T W L A F

73081 TGGGTCAACGAGCGCAACGAGGAGGAGCTGCGCGCGCGCGCATCCCTGCAACGCTC 73140
W V N Q R N D E V S L R G R R I P C T V

73141 AACCGGACACCACTCATCCACGCGCGCACCGCTACTACACTCCGCTCAGCACTG 73200
N P D H H L I H A A T A Y Y T S G Y D S

73201 GCGATGGCGGTGCCATGACCCCGCGGCTGCCGCGCTTCCGCGGCAAGGCGAGCG 73260
A M A V A I D P T G C R A F A G K G S R

73321	CTCTACCCCTTGCGCGGACCTCGACGCTGGTTCAGAGCCAACATCGGCTACTGCTAC L Y P L R R D D L D A W F N A N I G Y C T C Y	73320
73321	GTCCGCGCATGTGTTCCGCTCCAGCATCTGTCGCGCGGCAAGGTCACTGGCGCTCGCC V A D L M F G S S I V G G A A K V M G L A	73380
73381	CCCTACGGCAGACGCCCGCGCGCGCGCGGACGAGGAACCGCCGAGACCGCTGGCG P Y G R P A D G A G P D E E P P E T V R	73440
73441	GACTTCGCGCGCTGGTGGCCCTGGCCGACCGGCACCCGCGCCTCGTCGACGTGACGGC D F A A L V A L A D R H P R L V D V D G	73500
73501	AGGAAGCTCAACGCCCACTCGCCCACTCATCCAGCTGGGCGCTGAAGCGCAGCTGACC R K L N A T L A H Y I Q L G L E R Q L G T	73560
73561	GCGGTCTTCGCCAGCTCGCCCGCTGTGCGCCCGCAACGCGATCGACCCGGACATCTGC A V F A E L A P L C A R N G I A F D I C	73620
73621	CTCTCCGGGGTACGCCCTCAACGGCATCGCCACCACTCGCTTCGAGTGACGACGGC L S G G G T A L N A I A T Q L A F E S T G	73680
73681	TTGACGGCATGCACCTCCACCCGCGCTGGCGGACGACCGCAGCTCGCGCGGGC F E R M H L H A C G D D G T A I G A A	73740
73741	CTCTGGCACTGGCACCACTCTCTGGGCCACCCCGGCTCCACCAACCAACCGCGACCTC L W H W H H V L L G H P R L H H T N A D L	73800
73801	ATGTACTCCGTCGGTAGTACCCCGGACACCGTTCGGCGGGCGCTGGCGGGACACCG M Y S V R E Y P E H T V R R A V R D H C A	73860
73861	GCGGACCTCGTCTCGAGGAGCCGGGCACTACGTCGCGCAGGGCCGCGCAACTGTGTGCC A D L V V E E A C C G D Y V A R A E L V A	73920
73921	GGCGCGCGCTCATCGGCTGTGTACGACGGCGCGCGCGAGGTGGGCGCGGGCGCTGGGC G G A V A I G W Y D G A G E V G P R A L G	73980
73981	CACCGCAGCATCGTCGCGACCCGCGGACCCCGCCATCGCGGACCGGCTCAACTCCCG H R S I V A D P R D P A M R D R L N S Q	74040
74041	GTCAGTTCGCGGAACACTTCGCGCCCTTCGCGCGCTCGTGCTCAAGGACACGCGCG V K F R E H F R P F A P S V L L K E H A A	74100
74101	GAGTGGTTCGCGCTCTCCGACAGCCCTTCATGCTGCGGGCACCCGCTCTCAAGCCC E W F G L S D S P F M L R A T F V L K F	74160
74161	GGCGTCCCGCATCACCCAGCTCGACGGGAGCTCGAGATCCAGTCGGTCAACCGCGAG G V P A I T H V D G T S R I Q S V T R Q	74220
74221	GACACCCCGGCTTCCACGACCTCATCCAGCCTTCAAGGACCGTACGGGGATCCCCATG D T P A F H D L I H A F K D R T G I P M	74280
74281	GTGCTCAACACCGCCTCAACACCAAGGCGAGCGCATCGCGAGACACCGGAGGACGCC V L N T S L N T K G E P I A E T P E D A	74340
74341	CTGCGCACCTGCTCGGCTCCCGGCTCGAACACCTGGTGCTCCCGGGCTCATCGTCAGC L R T L L G S R L D H L V L P G L I V S	74400
74401	GGCCGAGCGCGCCCGCTCATGAGCGCCCGCGGGCGAGCGAGCCCGCGCCCGCGC G R T A A G R S *	74460 (orf10)
74461	TCGAACCGACATCGCGCGATCTGGGCGGAGACTCTCGGACGGGACAGCGTCGGCCCGC E R D I A A I W A E T L G R D S V G G P H	74520
74521	ACGAGGACTTTCGCGCGCTGGCGGCACTTCATCCGCGCATCAAGATCAACACCGG E D F A A L L G G N S I H A I K I T N R V	74580
74581	TGGAGGAACCTCGTCGACGCCGAGCTGTCCATCCGCGCTCTGCTCGAGACGCGCACCGTGG E E L V D A E L S I R V L L E T R T V A	74640
74641	CGCGCATGACGACCACTGACCGCACGCTCACGGGAGCGGAGACCGGTGAACACCGA G M T D H V H A T I L T G G E R D R *	74700 (orf9)

74701 CCTGCCCCGGCTGCTCGACCGGATCGCGGCTGCGCTCTCGTATCGCGGACGTCAT 74760
L P R L L D R I A G L R V L V I G D V I

74761 CCTCGACACTACGTCTGGGAGCCACCTCGGCGCTGTGCCGGAATCCCGCTCCCTGC 74820
L D T Y V W G A T S G L C R E S P V P A

74821 CGTCAACCTGACCTCCGTCGCCGCCAGTCGCGCGCGCCGCAACGTGCGCTGAACCT 74880
V T L T S V A H Q C G G A A N V A V N L

74881 CGGGCGCTCGCGCCGAACCGGTGCTCTCTCGCGACGGTGACGACCGCGCGCGCG 74940
R A L G A A E P V L L S A T G D D R A G R

74941 CGGGTCGCGGAAGCCCTCCGTGCGCGGACGTGACACCGCGCGGACTCTCTGACAGCC 75000
R L R E A L R A R D V D T G G L F V Q P

75001 CGCGCGGACACCGTCAACAAACGCGCGTCTATGGCGGACGACAGATGCTGCTCCGCT 75060
G R T T V T K R R V M A D G Q M L L R L

75061 CGACGAGGGCGCGGAACACCGCTTGCCCGTGGCGACGACACCGGAAGCGCGCTGCTCGA 75120
D E G G E H P L P V A T D T G S R L L E

75121 ACGGGCGCGCGCTGCTGCCCGCGTGAACCGGTGATGCTCTCCGATACGGGTACGG 75180
R A A G L L F A V D A V I V S D Y G Y A

75181 CGTGTGGGAGCCGACACCGTCCGCGGCTCGCGCACACCGGAACCTCGGCGCGTCCAC 75240
V W E P D T V A R L A A H R E L G P S T

75241 CTGTGTGTGACTCCCGCGCGCGCGCTTCAACCGGTGCGGGCCAGCGCGCTCAA 75300
L V V D S R R P A R F T A L R A S A V K

75301 ACCCAACACCGCGGAGGCGCTGCGCTGCTCGACGCGCGGAACCCCGCGCGCGCGC 75360
P N H A E A L R L L D A G E P P P G A

75361 CAGGGCGGACTGGGCGCGCGCTCGCGACCGGCTCTGCGCTGACGGGAGCGGAACG 75420
R A D W A A A L G D R L L R L T G A E R

75421 GGTGCGCTTCAACCTGGAGCGGACGATCACTGCTCTCGAACGGAACCGCGCGCGGT 75480
V A L T L D A D G S L L F E R D R P P V

75481 CGCACGTTCCCGCGGCGCGCGGCAACCGTCAACCGCGCGTCCGCGCGCGGACGCG 75540
R T F A R G S R A P V T A A V G A G D A

75541 CTTCACGCGCGCTCACCTCGCGCTCGCGCGCGCGACTCGCGGTGCGCGCGA 75600
F T A A L T L A L A A G A D S A V A A E

75601 ACTGGCTCCCGCGCGCGGACGCGCGTCCGCAACCCCGGACCAACCTGGCACGC 75660
L A S A A A G T A V A T P G T S T W H A

75661 CGACGAACCTGCGCGACTGCTCGCGGCAACCGCAAGGTCTGCGGACCGGACCGCTGC 75720
D E L R R L L G G T G K V C R T G T L P

75721 CGCCCGCTGCTCGACCGCGCGCGCGGACCGCGGTCTCTTCAACCAACGGCTGCT 75780
A R L L D P A A R D R R V V F T N G C F

75781 CGACCTCTGACGCGCGGCAACGTCTCTGCTGAGCGCGGCAAGGAACCTGGGCGACT 75840
D L L H G G H V S C L S R A K E L G D L

75841 GCTCGTGTGCGGTCAACTCCGACGCGAGCGTCCGACGCTCAAGGCGCCCGCTCGCC 75900
L V V G V N S D A S V R R L K G P R R P

75901 GGTGATCCCGCTCGCGAACGATGCGGTCTCGCGCGCTGAGCTGCTGGACCTGT 75960
V I P L A E R M R V L A A L S C V D L V

75961 GGTGCGCTTCAACGACGACGCGCGCGCGCTCATGAGGCGCTCGCGCGGAGGTCTA 76020
V P F D D D S P A A L I E A L R P E V Y

76021 GCGCAAGGGCGGGGACTACACCTCGCGACCTGCGGAAGCACCCCTCGTCCAGCGGT 76080
A K G G D Y T L A T L P E A P L V Q R L

76081 CGGCGCGTGTCTCCACTGCTCCCGAGCGTCCGCGACCTCCACCAACGACATATCG 76140
G G V V H L L P S V A D T S T T D I R

76141 GCGCATCCAGCCCTGTCCAGGACCGGAGAGACCCCATGAGCCACGCCATCGGA 76200
M S H A I G (orf8)

R I H A L S R T G E G D T P *

76201 CCGAGCCGGCTGATCCCCGCATCCGCGAAGCGCTCGGGGACGAGAAGGACCCCGCGCTC 76260
P S R L I P A I R E A L G D E K D P R L

76261 GCCCTCTACGTCCACGTCCCTCTTGTCTCTCTCAAGTGCCTTCTGCTACTGGGTACCC 76320
A L Y V H V P F C S S K C H F C D W V T

76321 GACATCCCGGTCCGACCGCTGCGCGGACAGCCGGGAACGTCTCGCCCTACGTCCACCGCC 76380
D I P V A R L R G D S R E R S P Y V T A

76381 CTCTGCGACCGATCCGCTTCTACGCGCCCGAGCTCACCGGCTCGGCTACCGCGCCCGAG 76440
L C D Q I R F Y G P Q L T R L G Y R P E

76441 GTCATGTACTGGGCGGCGGACCCCAACCGGCTCACCGGCGACGAGATGACGGCCGCTC 76500
V M Y W G G G T P T R L T G D E M T A V

76501 CACCAGGCCCTCGACGACGCTTCGACCTGACGGGACTCCGCCAGTGGTCCGTTGGAGAGC 76560
H Q A L D D A F D L T G L R Q W S V E S

76561 ACCCCGAACGACCTCGACCCCGCCACCTCTGACACCTCGCGCGCTCGGCGTCAACCGC 76620
T P N D L D P A T L D T L R G L G V T R

76621 GTCAGCGTCGGCGTCCAGTCCGCTCAACCGGTACAGCTGCGCAAGCGAGGCGCGGCCAC 76680
V S V G V Q S L N P Y Q L R K A G R A H

76681 TCGCGCGAACAGGCCCTGGCCCGCTCCCTCTGTCGCGCGCGCGCATCGACGAGTTC 76740
S R E Q A L A A V P L L R R A G I D E F

76741 AACGTCGACCTGATCGCGGCTTCCCGGCGAAGCCGTCGAGTCTTCGAGGAGACCCCTG 76800
N V D L I A G F P G E A V E S F E E T L

76801 CGCACCGCTCTCGCGCTCGACCCCGCGCAGCTCTCGCTTACCCCTACCGGACCAACCCC 76860
R T V L A L D P P H V S V Y P Y R A T P

76861 AAGACGCTCATGGCCATGCGCTCGACCGCGAGTTCGTCGAGGCCCGGAACCGGAGCGGC 76920
K T V M A M Q L D R E F V E A R N R D G

76921 ATGATCGACGCTTATGAACGGGCCATGGCGCGCTCGGCGCGCGCGCTATCACGAGTAC 76980
M I D A Y E R A M A A L G A A G Y H E Y

76981 TGCCACGGCTACTGGGTGCGGACGCGGCCACGAGGACGAGGACGCGCACTACAAGTAC 77040
C H G Y W V R D A R H E D Q D G N Y Y K Y

77041 GACCTGCGCGCGACAAGATCGGCTTTGGCAGCGCGCGGAATCGATCATCGTCAACAC 77100
D L A G D K I G F G S G A E S I I G H H

77101 CTGCTCTGGAACGAGAACAGCGCCTACGCGCGCTACCTGCTCGCCCCCGCGAGTTCTCC 77160
L L W N E N S A Y A R Y L L A P R E F T C

77161 GCGCGCCACCGGTTTACCAACCGCGAACCAGCGCTGACCGCGCGCGCTCGGCGCGCGG 77220
A A H R F T T A E P D R L T A P V G G A

77221 CTGATGACCGTGAAGGCGTGGTCTTTCGCGCGCTTCCGACAGCTGACCGGCTGAGCTTC 77280
L M T R E G V V F A R F R R L T G L D F

77281 GCGGACGTCCGCGCCACCGTACTTCCGCGAGTGGTTCGAGCTCCTGAGACGCTGCGGC 77340
A D V R A T P Y F R Q W F E L L E R C G

77341 GGCGCGCTTGTGAGAGCGCGTACAGCCTCCGCGCTGGAGCCGTCCACCATCCACCGCGC 77400
G R F V E T P Y S L R L E P S T I H R A

77401 TACATCACCCACTCGCCTACACCATGGCCATGGCCTGGCGCCCGGAAGCGCGCTGA 77457
Y I T H L A Y T M A H G L A P E R A *

SEQ ID NO: 2 ORFS BLM gene cluster ORFs 31-40

(notice this part is on the reverse strand and the last nucleotide (18660) is the first (1) on the whole cluster of 77457 bp. Also the last orf (40) is incomplete and contains frame shifts)

1 GTGACCGAGAACCTTCGTCGTGCCCGAATGCTCCAGCGGTACACCTATGAGATGGGT 60
M T E N L P S C P E C S S A Y T Y E M G
(orf31)

61 GCGCTCCGTTCTGCCCCGAATGCGGCCACGAGTGGCGCCCGCGACCGCGAGTCCGCG 120
A L L V C P E C G H E W P P A T A E S A

121 GACAAACCCGAGACGCGCGATCAGGGACGCGGTGCGCAACGTACTCGCGACGCGGAC 180
D N P E D G A I R D A V G N V L A D G D

181 ACCGTACGCTGGTCAAGAGCCTGAAGGTCAAGGCCACCCGACCGCGATCAAGGCGGGC 240
T V T V V K S L K V K G H P T G I K A G

241 ACCAAGGTGCGCAACATCCGCCTCGTGGAGGGTGTGGCGGCCACGACATCGACTGCAAG 300
T K V R N I R L V E G V A G H D I D C K

301 ATCGACGGGTTCGGCGCCATGCAGCTCAAGTCCAGCGTGGTCAAGAAGGTCTGACCGGTT 360
I D G F G A M Q L K S S V V K K V *

361 ACGCCGCGCCAGGCCCTGCCAGGCTCCACTACGCGCGCGCAACCGAGCCGGAACCGG 420

421 GCCCGGGCGCGCTCCAAGTCCCGTTCGCTGCGCGCGCGCGAGCGAGGCGTGTTCACC 480

481 CTGGGTGCGCGTCCCGTTCGTCACGCGTGTACACGCCACCAACGACGCGCACGGAATC 540

541 CCGGAACCTGCCACGTTCCCAAGTCCCGCGTCCCGGATCCCGCCGGACCGCGCTCGG 600

601 TCCGCGCGCGCGCGCGCGCGGTCGCCGCGCGCGGAGGGGGTCTCGCGCGCTG 660

661 GAACGCGCGCGGAAATTTACGTATAGGTAGAGATCCCGCGAAGCGATCGCGCGTTAT 720

721 GGCAGCATCCGCGCGCGCGCGCGCAGTTCCTCGGTCCCGGACGATGGCTCAAAAG 780

781 TGAGCGACGAAATCGCCGGATCGCGCGAGGACGTCGCGGGCCGACAGGACAAACCGG 840

841 GGATATATCAGCGCATTCACGAGTTCACGCGTTGACTGGAAATCGCCTACTTATCGCGTCA 900

901 CGCCTGTAGGATCATGCCCGGAATGGCCTCAGAGCTTTGAGTGCACCTACCTTGAGGTT 960
M A S D A L S A H L E V
(orf32)

961 TCCGACTGTGCGCAGCGCGGGGATCAGGTGACGAATGACGATCTGAATCGCGCGG 1020
S D C R Q R G G I T V T N D G S E L A G

1021 CAAACGTGGCGCGGTTCGCTTCGAGCGGTATTCGCGATCGCGCGGAGCGAACCC 1080
Q N V A A V R F E R Y S A I A P E R T A

1081 ATCCTGCACAAAGGTGCGCGACCGGTTACGACGAGCTCAACCGCGCGGACGATGACA 1140
I L H K G A A T G Y D E L N R R A E L T

1141 GCCACGCGCTGCGGACGCGGGCGCGGCCCTCGACCTGGTGGCGAGTGCCCTCCCA 1200
A T R L A D A G A G P S T L V A V A L P

1201 CGCGATCCGACCTCGTCGACCCCTGTGCGCCCTGCTCAAACCTGGGTGCGCATGCTT 1260
R D P D L V A T L C A L L K L G A A C L

1261 CCCCTGGATCCCGCATACCGCGGGCGGCTGCGGAGATCATGGCGGACGCGTCCCC 1320
P L D P G I P A G R L R E I M A D A S P

1321 GACGTTCTGTCACCAACCGTGGCGCTCGCTCGGCAATTCACCGGTGACGACCGCGCTT 1380
D V L V T T R A V A P A F T G D G P V L

1381 TTCTTGGACGACGCTCCTCCGACCTGCTCGGCCCTCTTCCACGGCACTCAGCGGGGACC 1440
F L D D A P P T C S A V L P R H S A G T

1441 GCGTCGGAATCGCCTATGTGCTGTAACCGACGACTCCTGACGAGAAGTCGAAAATTCG 1500
A S E I A Y V L Y P T T P D E K S E N S

1501 GTCTCTCTTATCGTGATATGGCGGCTACCTTGACGACCCCACTGCCGGGATTCCGGCG 1560
V V S Y R D M A R Y L E D P T A G I P A

1561 AGGGCGGAGATTCTCCGGCTGGTGGCGCCGCTCTCTGTCGGCGCTGCTGTCGTGCTGGAC 1620
R A E I L R L V A P L L S G G R L V L D

1621 GCCGACGAGACCGGCCCGGCGCGTACCCGTGAGGCGCCGCGACATGGTGAGGAC 1680
A D E T R P R P V T R E A P R D M V E D

1681 GTCTGGCGCAGGTCTGGTGGCCCTGCTCGCGTGACCGGTGGCGTGGCGGACCGC 1740
V V A Q V W C A V L G V D R V G V R D R

1741 TTCTTGACCTGGCGGCAAGTGGCTGGCGCGGTCCAGGTGGTGGCGCGCTCGGGAAG 1800
F F D L G G K S L A A V Q V V A R L R K

1801 CTGCTCGGCTCGAGCTGCGCTGGCGGCTCTTGACGCGCGCGAGCTCGAGGAGCTG 1860
L L G V E L P L R A L F D A P T V E E L

1861 GCGCCCGGGTGGCGGCCGAGCGCGCGCGCCAGGGCTCGGGGAGGCGCGCGCTC 1920
A A R V R A E Q A G G Q G V R E E A A L

1921 GAGCGGTGGCGCGAGCGAGCGCTGCGCTGTCTGTCGACAGCAACGCTGTGGTTC 1980
E P V G R S E P L P L S F A Q Q R L W F

1981 CTGGACCGTGTGATCCCGCGCGCTTCTACAGATGTGGACGCGCTTCCGCGTCCGG 2040
L D R L M P D R A F Y T M C D A F R V R

2041 GCGCGGATCGACCTGGGTGGCTGGCGCGGCCCTCGGATGCTGTGGACGCGACGAG 2100
G G I D L G A L R R A L R M L V G R H E

2101 ACGCTGCGGACGGCTTCTGCGAGCGGACGGTGTGCCGTACGAGCTGTGGTCCGGCC 2160
T L R T A F V E R D G V P Y Q L V G P A

2161 GACGGGCGCGTGGCGCGCGTGGCGCTCCACGCGGGTGCAGCTGTGCTGTGGAG 2220
D G P G A R R V A A P T R V D L S L L E

2221 CCGCCCGAGCGGAGGAGGCGGTGGCGAACCTGGTGGCGCGGAGCGCGGACCCGCTTC 2280
P A E R E E A V R N L V A A E A R T P F

2281 CGGCGCGGACGCGCGCTGCTGCGCGTGGTGGCGCGCTGCGGACGATGATCAC 2340
R P A D G A L L R V V V A R L A D D D H

2341 GTGCTGGTGTGACGACGACCATGCTGCTCCGACGCGTGTCCGTGGTGTGTGCTGTG 2400
V L V V S T H H I V S D A W S V G V L V

2401 GACGAACCTGACGCGCTGTACCGCGAGTGGCTCACCGGAGATCCCGCGCGCTGCCCCG 2460
D E L G R L Y R E C V T G D P A A L P F

2461 CCGGCGTCCAGTACGCGGACTTCCGCGTCTGGCAGCGGCGCTGGATGGCGGTCGGTG 2520
P A V Q Y A D F A V W Q R A W M A G P V

2521 CAGGAGGAGCATCTCGCTACTGGAAGCGGCGCTTGGACGGCGCTCCCTGGTGTGCGG 2580
Q E E H L A Y W K R A L D G A P S V L R

2581 CTGCCATGGACCCGCGCGCGCTGCGAGTCCGAGCGGCGGAGACGTCGGGTTC 2640

L P M D H P R P A V Q S E R G E T V G F

2641 GCGCTGCCCGACGCGCTGGTCGCGCGCTGGAGAAGCTGGGCCGGAGCAGGCGGCCAAC 2700
A L P D A L V A A L E K L G R E Q G A T

2701 CTGTTTCATGACGTGCTCGGCGCTTCCAGGCTCTGCTGGCGGTCACGCGGGCAGAG 2750
L F M T L L L G A F Q V L L A R H A G Q E

2761 GACATCGTGGTCGGCGTGCCTGGCGGGGCGCACCCGGAGCAGAGACGGAACCTCTGGTC 2820
D I V V G V P A A G R T R T E T E P L V

2821 GCGTTCTCTCAACACGCTTCCCTTGGCGGCGATCTGCGCTCGGCGGCTGCTGCTCCGG 2880
G F P V N T L P L R A I C A P G L S F R

2881 GACCTGCTGGACAGGTGGCGAGGCGCGCTCGGCGGCTTGGCCCATCAGGACCTCCCG 2940
D L L D Q V R E A A L G A F A H Q D L P

2941 TTCAGGCGCTGGTCGAGGCGCTCGCACCGAGCGGACCTCGGCCACAATCCCTCGTC 3000
F E A L V E A L A P E R D L G H N P L V

3001 CAGGTCACTTCCAGCTCCTGGGACACCGCGCGCGCGGACCTGATCGGACCGAG 3060
Q V T F Q L L G T P A A R P D L I G T E

3061 GTGAGCGGTACCGGTCAGGAGGCGCTCTCGAGTTCGACCTGCTCCGTGACATCAG 3120
V E R Y P V Q E A V S Q F D L S L D I K

3121 CCGGCGGACGAGCTTCTACCGGGGATCCTGAACCTACTGCCCCGACCTGTGACCGGA 3180
R A D D G S Y R G I L N Y C P D L P D R

3181 CCGCGATGGAGGTCTGGTCGGGCACTACTGACGCTGCTCGCGCGCGCGCGCGGAC 3240
R R M E V L V G H Y L T L L G A A A A D

3241 CCGGCGCGCGGATCGGTGAGCTGCGCTGTGCGACGGGCGGAGCGGCTGCGGCTGCTC 3300
P G R P I G E L P L S D G A E R L R L L

3301 GACGGGTTCGGGAAGCGGACGCGGCTACCGGGCGGGAAGCTTTCGGAGCGGTTC 3360
D G F G K R D A A Y A G P G S V P E R F

3361 GCGAGGTGGCGCGGACGCGACCGGACGCGGCGGCTGACCTGTGCGCGACACGCTC 3420
A E V A R T A P D A R A V T C G A T T L

3421 ACGTTCGCGAGCTGAACGACCGGCTGGAGCGCTGGCACAGGCACTGCTCGGCGCGGG 3480
T F A E L N D R V E R L A Q A L L G A G

3481 GTCACCGCGGAGACGCGGTCGCGGTCGCGCTGCCCGTTCACCGACAGCTGCTGCGC 3540
V T R E T P V A V R L P R S T D S V V A

3541 CTGCTGGCGCTATGCGGCGGCGGCGCTACGTCCCCCTGACACCGGACCTGGCGCGG 3600
L L A V M R A G G V Y V P L D P D W P A

3601 GACCGCACCGCTTACATCTGGACGACCGGCGGCTCGCTGCTCATCACCGGACGCTG 3660
R T A Y I L D D T A A S V V I T R D L

3661 CCGGCACTCCCGGTCGCTCCACGTGACCGCGCGCGGCGCGGCGGCGGCGGCTGCTA 3720
P A L P G R L H V D P R R P A A D G L V

3721 CCGGCGCGCGCATGACCGGATCAGGCGGCGTACGTCTACACGTCGCGCTGACG 3780
P A P R I D P D Q A A Y V I Y T S G S T

3781 GCGCGCGGAGGCGCTGCTGCTCGGCACTGCTGCTGACCACTCACCAGCGCGCTG 3840
G A P K G V V V R H R S L N H L T S A L

3841 CAGGCCACCTTTCTCGGCCACGACCGGTATCTCGCCGGGCGCCGCGGTACCCGCCGGG 3900
Q A T F L G H D P Y L A G A D G V P P G

3901 GACGCGAAGCTGCGTACGACGCTCACCGCGCCCTTCAGTTTCGACGCGTCCATGGAGCAA 3960
D A K L R R T T L T A P F T F D A S M E Q

3961 CTGAGCTGGAATGCTGGCCGCTCAGAGCTGTTTCATCGTCCCGAGGACGTGCGGCGCGAC 4020
L S W M L A G H E L F I V P E D V R R D

4021 CCTCGGCGCTGGTCCGTTCTCGCGGAGCACCGGATCGACGTCTAGACGACGACCTCC 4080
P S A L V R F V R E H R I D V I D T T S

4081 TCGCAGCTCGAACTCCTCGTATCGCACGGGCTGTTGAGACGAGAGTGGGCGCGCTCCATG 4140
S Q L E L L V S H G L L D G E W A P S M

4141 GTCATGGTGGGTGGCGAGGCGGTCTCGCCGTGCTGGGCGGACCTTGCGGACCGAGCGG 4200
V M V G G E A V S P S L W R T L R D Q R

4201 CGCACTCGCTGTTTCAACCTGTACGGGCTACGGAGGCGACGGTTCGACGCCACTGCGCAC 4260
R T R C F N L Y G P T E A T V D A T C H

4261 GACCTGTCCGACCCCGCGACGTCCCGCTCATCGGACCCCACTCCCGGACACCCACGCTC 4320
D L S D P A D V P V I G T P L P H T H V

4321 CGCGTCTCGACGACCGACTGCGACCCGTACCGCTGGGCGTCCCGCGAGATCTACCTC 4380
R V L D D R L R F V P V G V A G E I Y L

4381 GGCGGAACCGGCTGGCCCGGGCTACCTCAACCGCCCGCCCTTCACGCGCGACGCTTC 4440
G G T G L A R G Y L N R P A L T A R R F

4441 GTGCGCGACCCCTACCCCGACACCCCGGCGAGCGGCTGTACCGACCGGACGACGCGCC 4500
V A D P Y P D T P G S R L Y R T G D R A

4501 CGCTGGCGCCCGGACGCGACCTTCGAATACCTGGGACGACCGACGACCAATCAAGATC 4560
R W R F D G T L E Y L G R T D D Q I K I

4561 CGCGGCTTCGCGTCAACCGGCGGAAATCGAGGCGCTCCTCACCCACGACCCGCGCTC 4620
R G F R V S P G S I E A V L T H H P A V

4621 AAGGAAGCGCGCTGCTCGACGCGGACGCGCGGCTGGTGGCTCACTGACGCTCAAGCTCGG 4680
K E A A V V D D A H A R L V A Y V T L A

4681 GAAGCGGCGCGCGCGGCCCGACCGACGTACCGCGGTTGCGGCGAGGCGGCGTCCCGCC 4740
E G G G A G P T D V R R F A Q G R L P A

4741 CACATGGTGCCTGCGCGTGGTCTGCTCGAGGCGGCTGCCACTGACGCTCGAACGGAAG 4800
H M V P S A V V V L E A L P L T S N G K

4801 CTGGACGCGCGCGCTGCGCGGCGCGCGCGGCGGACCGGAACTGGATGCTCGCTTC 4860
L D R A R L P A P A A G R P E L D V R F

4861 GTGGCGCGCGCGACATGGGTGAGGAGGTCTGGCGCAGGTCTGGTGCGCCGTCTGGAC 4920
V A F R D M V E E V V A Q V W C A V L G

4921 GTCGACCGGCTCGGTGTGCAACGACGACTTCTCGAGCTGGGCGGGCACTGTTGCTGTG 4980
V D R V G V H D D F F E L G G H S L L V

4981 GTCCAGGTGATGACCGGATACGAAGCTGCTCGGCGTTCGAGGTGCGGTGCGGGAGCTG 5040
V Q V M T R I R K L L G V E V P L R E L

5041 TTCGACCGCGCGACGCTCGAGGAGCTCGCCCGCGCTCGCGCGGACGACGCGGAGGC 5100

F D A A T V E E L A A R V R A A R T E S G

5101 CTCGGCCGGGGGCGCCCGCCCTCGGGCCGGTGAGACGGAGCGGGCGCTGCAGCTG 5160
L G R G A A F P L G F V D R S G F L F L L

5161 TCGTTGCGCAGCAACGCCCTTTGGTACCTCGATCAGTTGGCGCCCGACAGTGTCTCCTAC 5220
S F A Q Q R R L W Y L D Q L A P D S V S Y

5221 AACATGTGCGACGCTACCGGGTCCGCGGCCCTCTCGACCTGGACGCGCTGCGGCGGGCG 5280
N M C D A Y R V R G P L D L D A L R R A

5281 CTGCGGAGCGCTGGTGGAGCGGCAGAGACGCTGCGGACGGCGTTCGTGAGCGGGACGGG 5340
L R T L V E R H E T L R T A F V E R D G

5341 GTGCCCCACCGAGTGGTCTCGGCGCCCGACGCGCGCGCGCGCGCGCGCGCGAGGTC 5400
V P H Q V V S A P D A P A A R R A A E V

5401 GTGCGGATCGAGGCGCGCGGCGGACCGACGAGGCGGTGGGACCTGGTGGCGCGGAG 5460
V R I E A A G R T D E A V R D L V A A E

5461 GCGCGCACCCCGTTCCGGCGGCGGACGCGCGCGTGAITCGCGTGGCGGTGCCCGCTG 5520
A R T P F R P A D G A L M R V A V A R L

5521 GCGGACGACGATCAGTGGTGGTGGTACACGACACACATCGTCTCGGACGGCTGGTGG 5580
A D D D H V L V V T T H H I V S D G W S

5581 GTCGACATCCTGGTGACGAATTGGGGCGCTGTACCGGGAACACGCTCAGGGTGACCC 5640
V D I L V D E L G R L Y R E H V T G D P

5641 GCGGGGCTCCTCCGCTCGACCTCAGTACGCCGACTTCGCCGCTGCGAGCGGTCTCTGG 5700
A G L P P L D V Q Y A D F A V W Q R S N

5701 ATGACCGGCCCCGTGGGGAGGAGCACTCGCGTACTGGAAGCGGGCCCTGGACGGGGCA 5760
M T G P V R E E H L A Y W K R A L D G A

5761 CCTCGGTCCTGCGGCTGCCGCGGACCATCGCGTCCCGCGCTCAGTCCGAGCGGGC 5820
P S V L R L P A D H P R P A V Q S Q R G

5821 GAGACGCTCAGTTCCCGCTGCCGACCACTGGTCGCGGCGCTGGAAGCGCTCTGCGG 5880
E T V E F P L P A P L V A R L E A L C R

5881 GAGCAGGGCGTCAACCTGTTTCATGGCGCTCTTCGCGCGGTTCCAGGTGTTGTCGCGCG 5940
E Q G V T L F P M A L F G A F Q V L L A R

5941 TACAGCGGTGAGGACGAGTGGTGGGCGTGGCGGCGGAACGACCGCGCGGAG 6000
Y S G Q D D V V V G V P T A N R T R A E

6001 ACCGAGCCCCGTGGTGGCTTCTCGTCAACACCTTCGCGTACGGGTCCGCTGCTCGCG 6060
T E P L V G F P V N T L P V R V A C S P

6061 GAGCTGTGTTCCGCGCCCTGCTGACCGGGTCCGCGAGGCGCGGCTGGGCGCCTTCGCC 6120
L S F R A L L D R V R E A A L G A F A

6121 CATCAGGACCTGCCCTTCGAGGCGCTGGTGGGCGCTCGCGCCCGAGCGGACCTGGGC 6180
H Q D L P F E A L V E A L A P E R D L G

6181 CACCACCTCTCGTGCAGGTACCTTCCAACTCTCGACGCTCCCGACGAGAGGCTCGTC 6240
H H P L V Q V T F Q L L D A P D E R L V

6241 CTGACGCGCAGGACTGCGTCTCGCTCGGCTTCGGCGGTGTGACGAGCGGTTCGACCTG 6300
L H G T D C V S L G F G G V T S R F D L

6301 TCCCTCGACGTCGTCTCGGGGCGGGGGGAAGCGGTGCGTCTGACGTACTGTCGCGAC 6360
S L D V V S G R R G K R C V L T Y C P D

6361 CTGTTGACACGGGCCCCGATGGAGGTGCTGGCCGGCCACTACCTGACCCCTGCTCGGCGG 6420
L F D R P R M E V L A G H Y L T L L G A

6421 GCGGCGGACGATCCCGGTCTCCGCGTCGCGGACCTCCCGCTGAGCGACGACGTCGAAAGG 6480
A A D D P G L R V G D L P L S D D V E R

6481 CTGCGCTGCTGGGGGGTCCGCGCGCGGTACCTGCCCGCGCCCGGGGCGAGACGCTC 6540
L R L L G G S R P R Y L P A P G A E T V

6541 CTTGACGCCTTCGCGCGCAGGTGCGGGCGACACCGACGCGCCCGCGCTGCTCAACGG 6600
P D A F A A Q V R A T P D A P A L V H G

6601 GACTCGACGCTGACGTTGCGCGAGCTGGACACCGGGTCAACCGCCCTGGCGCTGCGGTTG 6660
D S T L T F A E L D T R V T A L A V R L

6661 CGCGCTGCGGGGTGGCGCGGACGCGGTGCGGGTGTGCTGCGCGCTCCGCGGAC 6720
R R C G V A A E T P V A V C L P R S A D

6721 GCGCTGTGGCCCTCTCGCGGCTCTGCGGGCGGGCGGCGTCTATGTCCAGTGAATCG 6780
A V V A L L A V L R A G G V Y V P V D P

6781 GAGTGGCCCTCCGGCGCGGTGCGCCACGTCCTCGACGAGAACCGCGCGCCCGTGTGTCATC 6840
E W P S G R V A H V L D E T A A P V V I

6841 ACCCGGACCTGCCCGCGATCCCGCGCGCTCCACCTCGACCGCGCGCGCGCGCGCC 6900
T R D L P A D P G R V H L D P R Q A P A

6901 GAGGACCGGATCCCGTGCAGCGCTCCACCGGACGAGCGCGGTACATCATCTTCAAC 6960
D D R D P L P R L H R D Q A A Y I I F T

6961 TCGGCTCCACCGCGCGCCCAAGGGCGTGTGTGCGACAGGCTCCCTGTATACACTC 7020
S G S T G A P K G V V V R H G S L Y H L

7021 CTGGGCCACGTACGGCGCATGGCGGAGGGCGGCGCGCGGAAACGTCGCGCACACACC 7080
L G H V R R M A E G G P R R N V A H T T

7081 GCGATGACCTTCGACCGTGTGTTGGAACAGTTCCTGTGGTCTGTGCGCGGACACACCTG 7140
A M T F D P S L E Q F L W L V A G H T L

7141 CAGTTCGCGCGGAGGAGTGGCGCGATCCCGAGGCGTGGTGGCCCTGTGTGCGGCG 7200
H V A P E E V R R D P E A L V A L V R R

7201 GCGCGATGACGTCCTCAACGTCAACCGCTCCACCTGACCTGTGTATGAGGCGCGG 7260
A A I D V L N V T P S H L T L L I E A G

7261 CTGTCGAGGGCGACCGGGTCCGGGTACGTCCTGTGTGTGGCGAGGCGGTGCGCGG 7320
L L E G D R V P G T V L V G G E A V P A

7321 GCGCTGTGGCGGACCTGCGCGAAGCGGAGCGAGCCACCGCTCTTCAACCTGTACGCG 7380
A L W R T L L R E R T G A T R F P N L Y G

7381 CCTACGGAGGCGAGGTGACGCGCACCTGCCACGACCTGTCCGACCGCGCGAGTCCCG 7440
P T E A T V D A T C H D L S D P A D V P

7441 GTCATCGGACCCCACTCCCAACACCAAGTCGCGGTGCTGACGACGACGACGACGAC 7500
V I G T P L P H T H V R V L D D R L R P

7501 GTACCGTGGGCGTCGCGCGGAAATCTACTCGCGGAAACGCGCTGGCCCGGGCTAC 7560

V P V G V A G E I Y L G G T G L A R G Y

7561 CTCACCGCCCGCCCTCACCGCCCAACGCTTCGTCGCGACCCCTACCCGACACCC 7620
L N R P A L T A Q R F V A D P Y P D T F

7621 GGCAGCCGCTGTACCGCACCGGCGACCGCGCCGCTGGCGCCCGACGGCACCTCGAA 7680
G S R L Y R T G D R A R W R P D G T L E

7681 TACCTGGGACGCACCGACGACCAATCAAGATCGCGGCTTCGCGCTGAACCCGCGAG 7740
Y L G R T D D Q I K I R G F R V E P G E

7741 ATCGAAGCCGTCTCTACCCACACCCCGCGTCAAGGAAGCGCGCTCACCGTGGCGACC 7800
I E A V L T H H P A V K E A A V T V A T

7801 GACGACGGTCCCGCCGCTGGTCGCCCTCGTCGCCCGCCCGCGCGCCCGCACGGC 7860
D D G A A R L V A L V V P A P R A P H G

7861 GATTGGCCGACGCGCCCGGACGCGCCAGGTGAGGAGTGAACCGCTCTTGAGGCG 7920
D S A D G A P D A Q V E E W N A V F E A

7921 ACCCACCGACGCGCGCGCGGCGAATCACCTTCAACATCAAGGGCTGGAACGACAGC 7980
T H T D A A D G E L T F N I K G W N D S

7981 CTCACCGGTGCGCGATCCCCCGCAACATGCGGGAATGGGTGACACCACTCGCC 8040
T G A P I P A E H M R E W V D T T V A

8041 CGGCTCTTGAACGCGCGCGGCGCGTCTGAGAGATCGGCGAGTGGCACCGGCTGCTG 8100
R L L E R P A E R V L E I G S G T G L L

8101 ATGTGGCGGCTGCTGCGCGACGTACCGAGTACACCGAAGCGACTTCTCGCGCGCGCC 8160
M W R L L P H V T E Y T G T D F S R F A

8161 GTGGACTGGCTCCGGGACGGGCTGCGCGCCCGCCCGCGACCGGATACGGCTGCTGAC 8220
V D W L R D G L R R R R P A H R V R L L H

8221 CGGAGGCGACCGACTTCACCGGCGTCCGCGCGCTCCACGACCTCGTCTGTCAAC 8280
R E A T D F T G V R A A S T D L V V V N

8281 TCGTCTGTCCAGTACTTCCCGACCGCGCTTACCTCGACACCGCTCTGCGCGCGCCCT 8340
S V V Q Y F P D R A Y L D T F V L A R A L

8341 GACGCCACGCGCGACCGAGGGCGCGTCTCGTGGGCGAGTGGCGAACCTGGCGCTCGCC 8400
D A T A D R G R V F V G D V R N L A L A

8401 CGCGAGTTCTACGCGCGTACGCGCTCGCCCAACGCGCTCGGGCGCGCGCGCGGAC 8460
P Q F Y A R Q A L A H A G P G A A A R D

8461 GTGGCGCGCGCGCGCGAGTTGCGCGGATGAGCGAGTGTGCGTGTCTCCCGCG 8520
V A R A A G E F A A M D G E L L V S P A

8521 TACTTCGCGCGCTCGCGCGCGCTCGCGCCCGTACCGGCGTCTGAGATCTGCGCGCGC 8580
Y F A A L A A R S P R V T G V E I L P R

8581 CGGGACGCGACCGCAACGAGATGAGCCTGTACGCTACGACGTGTGTCTGACGTGGC 8640
R G R H R N E M S L Y R Y D V V L H V G

8641 GGTACGCGCGCGCGCGCGGAGGTGCTCACCTGGGGCGACAGGTGACGAC 8700
G D R P A A P E A E V L T W G D Q V H D

8701 CTCGCTGCTGTCTCGCGCGCGCTCGCGCGCGGGCGCGACGCGCTCTGCTGCGCGCG 8760
L A S L S A R L G R G G P D A L L V R G

8761 GTGCCCAACGACCGTCTGAAGCGGGACAAACGAGCTGCTGACGACCCGCGCCGACGAG 8820
V A N D R L T R D N E L L D A P A R T T

8821 GCCGTCGAGCCCGAGGACCTGTGGGGGCTGGCGGACTCCACCCCTACCGGGTGAGCGTC 8880
A V E P E D L W G L A D S T P Y R V S V

8881 AGCTGGCGCGCCCGCATCCGCGGGGCGCATGGACGTCCTGCTGCTCGGCGGGACGCC 8940
S W A A A A D P R G A M D V L L V R R D A

8941 CACGACGAGGTCCGCTGCTGCTCCCCCAACCCCTACCGGAGCCCTCGGCACCGCTGAG 9000
H D D G P L L V P H P V P E P S A P L T

9001 AACACGCGGACCCGCGACCCGTCGCGGGGCAAGGGGCTCGGCCGCGACGGGCTGCGT 9060
N T P T R H P S A R Q G G S A A D G L R

9061 TCCTGGCTGCGCGAGGCGCTTCCCGCGCACCTGCTGCCCGCGAGGATCACCGAGGTGAC 9120
S W L A E R L P A H L L P A R I T E V D

9121 GCGCTGCCCGCGACCGGCAACGCGCAAGCTCGACCGGGGCGCGCTCGCGCGGACTCGTGAC 9180
A L P R T G T G K L D R G A L G G L V T

9181 GCGGCGGTGGCGCCCGGGCGGGCGACCGCCCGCCACCGCCCGGCTACGGGTCTCGAA 9240
A G R G A R A A G D R P A T A P R T G L E

9241 CGGACCCCTGGCGGACCGCTGGGCGGGGTGCTCGGCTTCGCGAGTCGGCGTGACAGAG 9300
R T L A D A W A R V L G L P E V G H E

9301 AACTTCTTCGCCCTCGGCGGCGACTCCCTCCTCGCGCTCAGGGCTGTGCGCCGGTGGCG 9360
N F F A L G G D S L L A V R A V A R C R

9361 CGTGCGGGGTCCGACTGACCTCCGGCAGTTGCTGAGCGAGCAGACCTGCGCGCGCTC 9420
R A G V R L T V R Q L L S E Q T V T A A L

9421 GCGCGGCCCTCGAGGAGGAGTCTCAATGATGAAGTCAAGCGGCTTGGCGGACCGGCGAG 9480
A A A L E E E S Q M M K S S R L R D R Q L
(orf33)

9481 TCGGGGTGAAGACCGGTTGTGCGCGAGGAGGCCACAGGACGCTGGCGCGGACGCGCT 9540
G G E D P V V A Q E S P Q D A G P T P C

9541 GCCAGGCGATGACGGCTTGAACGTGTTTGACGCGCTCGCGCGCTTCTTGAGGTAGAAG 9600
Q G D D G L N V F A A L A A L L E V E V

9601 TCCCGGTTCGGCCCTCCCGCATCATGTGGTTTGGCGGACATGTAGAACTACTGTCGC 9660
P V R P L P H H A G L G R H V E H S S Q

9661 AGGCGGCGGTGTAGCGCTTGGGCGGATGCAAGTTGCCAGTGCAGACCGAGCTGCGCG 9720
A A A V A L G F M Q V A S A T T G V A G

9721 GGGACGGGACCCAGGCGGCGCGGAGGCCAGGTGACCGCGCTCGCGGTAGGCGGTGAGG 9780
D G H Q A G R R G Q V T G V G V G R E V

9781 TCGCGGCGGCGACGAACTCGGCGCCGAGGATCGGCCCATGCGCCGCGAGACTCG 9840
A G G D D E L G A E D R P H A R Q R L D

9841 ATGATCTCGGCTGTGGATGGCTGCGAAGCTCTCGCGGATCTGCTGGTCAATCCGCTTC 9900
D L G L W M A A E R L A D L L V N P L Q

9901 AGACGGTCGTCCAGGGCCAGGATCTGCGCGGCCAGGTGAGCCACGATCTGGGCGGCGAG 9960
T V V Q G Q D L R G Q V S H D L G G D V

9961 TCCTCCCGGGCAGCGCGTCTGCTGAGCCTGGCAGCCTCCAGCGCGTTCGCGGAGC 10020
L P G Q R G L L S L G S L Q R R R G D G

10021 GCGTCGGCACGCGCACGCTCGGTGGCCAGCCAGGCGTTCAGCGGGGCCGCGCGG 10080
V G T A H A S V G Q P G R Q P G F A A A

10081 CGGCGGAGAGCTGCGGGGCTGCTGGTAGCCGTCAGCAGGACCAAGCGCGCTTCTGCGAG 10140
A E S C R G L V A R Q Q D Q R A L L R A

10141 CTGTAGTCGAAGGCCGCTTCAGCGCGGGGAAGACGCGTTCAGCGTCTGCGGAGCAGG 10200
V V E G P P F Q R G E D A G Q R V A E T V

10201 TTGATCATCTGACCCGGTCGGCCACGAGGTGCGAAGCTGGGCGCTCAGCAGCGCAGG 10260
D H P D P V G H E V G T V G G Q Q R E V

10261 TCGGCGGCCAGCTGGGCGGCGACGTGATGACCGGAAGTCCCGTGGTTCGGGCGGTT 10320
G G Q L G G H V D R R E V P S V A G G F

10321 TCGGCGAGTACGTAGGCGTTCGGGCGTCTGCTTCGCTTCGCCCCGTAAAGCGCGGAC 10380
G D D V G V A G V G L R L A P V S A G H

10381 ATGCGGTGACCGTGGCGCGGGCACTAGACGCGCTGCGCGTGGGCGCGGAGCAGG 10440
A V D R A A G H V D G L L A V G R E Q G

10441 GCCAGCAGCAGCGCGGAGGACGTGCGGAGATGTCCTGCGCCAGTGAACCTGTCGCGC 10500
Q Q Q R G G R A G D V H C P V D L V G Q

10501 AGGTGAGGATCTACCCATGGCGGTGAGGATCGCGACTCATGTTGCGATCTTCTTC 10560
V E D L T H G G Q D R R L I V A D L L R

10561 GACCACAGCGTCACACCGGTCTCGTGCACACCGCGCCACTGATGCGCCCTTGCCTCGG 10620
P Q R H T G L V D H R R F V M P L A R V

10621 TCGATCCCGGCCAGCACCAGGCGCGCGTCTGCGCCACTGCGCCCTCTCACTCGGAACA 10680
D P G P D P G P S L A H S P L L T P N S

10681 GCATCCGTCGACCCGAGGAACACCCGCTGTCTCTCGTAAAGAGCAGCAGGAGCGCA 10740
I P S T R G T P R C H L R K K R P K R T

10741 CATCTCAATCAGCAGCGAGGCGCCCGGAGAACCGGCGGCACTCTTGTAAAGCACT 10800
S Q S A A R A P R R T G R P L L V S H *

10801 GACGGCAGAGAACCATAAGCCACACCGCGCCCTCCCGGCGCGCTAACAACTTACGGAGA 10860

10861 ACCATGACTGACCTGCGGTGGGTACCGTGCACCTCACCGGTGAGGAGAGCGCGGAGTCT 10920
M T D L P L R T V A L T G E E S A E V
(orf34)

10921 GACGACTGCTGCGCAGCTGCGCGACGTCGCGGTGACCTCCACCTGGGAGCTGCTGCAC 10980
D D L L R T L A D V P V D S T V G L L H

10981 OGCAACCGGCTGCGCGCACAGGAAGTCCGCTGCGCATCGCGCGGAGCTCACGGGGATG 11040
R T R L A A Q E L P L R I R A E L T G M

11041 CGGCTCTACGACAGCCCGCGCGCCCTCGTGTACGGGCTTCGGCGCTGAAGCAAGG 11100
R L Y D S P R A L V V T G F G V D D E R

11101 ATCGGACGACCCCCCGCGCGCGTCCCGCGCGGATCCCGAGCGGACCGCGCTGAGAG 11160
I G P T P A A R P A P D P E R T R D L E

11161 CTGCTGCTTTTGTGCAAGCGGCGCTCGGCGAGGCGTTCGGCTGGGCGACCCAGCAG 11220
L L L L L H A A L L G E A F G W A T Q Q

11221 AACGCCGGCTCGTCCACGACGCTGCTGCCGTTCCCGGTGAGGAGACCGCGCAGATGGT 11280
N G R L V H D V L P V P G E E T A Q M G

11281 TCAGCAGCGAGACCGAGCTGCTGTGGCACCCGAGGACCGCTTCCACCGCTGCGCTGC 11340
S S S E T E L L W H T R D A F H P L R C

11341 GACTACGTGGCGCTGCTGTGCTGCGCAACCCAGCGCGCGACCACTGGGCTGG 11400
D Y V G L L C L R N H Q R A A T T V G W

11401 CCGACCTGTTCCCGGCTCACACCGAGGACCGTGCCTGCTCTCTCGAACCCCGTATCTG 11460
P D L S R L T T E D R A V L L E P R Y L

11461 ATCCGCCCGACACTCGCACACGCCCGCGCAGAACCGGACCGGCGCTCGCGCGAG 11520
I R P D T S H T P A Q N A T G T R S A E

11521 CGTTTCGCGCGATCGCCGAGATGGACGACGCCCGCGAGCGCTGCGCTCTGTTCGGC 11580
R F A A I A E M D D A P E R V A V L F G

11581 GACCCCGAGGACCGTACTCTGCGGATCGACCCGCTACATGAGCCCGGCGCGCGGAC 11640
D P E D P Y L R I D P A Y M S P A P G D

11641 GCGGCCCGCGCGCGCTACGACACCGTCAACCGCTCATCGAGGACGATGCGCGAC 11700
A A A R R A Y D T V T A L I E D E L R H

11701 GTGCTCTGGACGCGCGTTCACTGCTGTGTGACAACTACCAAGCGGTGACGCGCGC 11760
V V L D A G S L L L V D N Y Q A V H G R

11761 AAGCGTTGCGCGCGCTACGACGCGCGGACCGCTGCTCAACCGCTCAACATCAC 11820
K P F A A A Y D G R D R W L K R V N I T

11821 CCGACCTGCGCGTTCCCGGTGCGCGCGCGCTCGGCGCTGCTGTGTGAGGG 11880
R D L R R S R S A R R S A T S L L V *

11881 AGGCACCATGGATTTCCTCTCAACCGCTCAACCGCTGTTGAGCGCGCTGCGACGG 11940
M D P P L T R V N P W F S G G C D G
(orf35)

11941 CCGCCCCCGGTGCGGCTGTGCGCGCTGCCGTACGCGGCGGACCGCGCGCTCTCAA 12000
R P R V R L C A L P Y A G G T A A V F K

12001 GGACTGGCCCGCGCTGCCCCCGAGTGGAGCTGCTCACCGCGCACCTGCGCGGACG 12060
D W P A A L P P G V E L L T A H L P G R

12061 CCGCGACCGTTACCGAACCGCCCCCGGCCACCTTGAGGAGACCGCGAGCGCTGTG 12120
G D R F T E P P P A T L E E T A E R L C

12121 CGAGGCGCTGCGCGCGAGTGACCTGCCACCGTGTCTCGGCCACAGCATGGCGCCCT 12180
E A L P P S D L P T V V L G H S M G A L

12181 GCTGGGTACGAAGTGGCGGCGGCTGCGGCGCGGCGCGCGCCCCCACTGCTGAT 12240
L G Y E V A A R L A A R G R A P N L I

12241 CGCCGCGGCTGCGCTCCCCCGACGTTCCGCGGAGCGCTTCCGCTCGGTGACCGAGGC 12300
A A A C R P P H V P P D A S G P V T E A

12301 CGAGCTGGCGGCCACCTGCGGGCGGAACGCCCATGGAGACAGCGCCCTGAGGACGAGGA 12360
E L A A T L R A E R P W D T A L R D E E

12361 ACTGATGGAAGCGGTGCTGCCGCCCTGCTGCGCGACATCACGCGCGGCGACCGCTACCA 12420
L M E A V L P A L V A D I T A G D R Y H

12421 CCGCCGCGGCCCCCGCTGACCTCCCGCTGAAGTCTACATCGGCGCCGACGACGA 12480

R P R P R P L D L P L K V Y I G A D D D

12481 CGGCACCGACTGGCGCACACCTGGGCTGGCGCGTGCACCGCCCGGACTGCGAGGT 12540
G T D W R T T L G W R A C T A R D C B V

12541 CGTGTCTCTGCCGCGGCCTACTTCTTGGAGACCGACCGCGCGCTCTCACCOC 12600
V V L P G G H Y F L E T D R A A V L T R

12601 CGTGCACCGACCTCGCCGAAGCCGAGGTAGGGGCATGACCGCGCGCTCGACGCCACA 12660
V A T D L A E A E V G A *
M T A R V D A T
(orf36)

12661 CCCACCTACTTGGCGGTGCTGGCGGTGGCGAGGCCCGCCCGCTCTCGGCAGCTGC 12720
P T Y L A V L A V R E A R A P L L G S C

12721 CTGGCCCGCATGTCTCTCGGCTGCTGCGCTCGCCCTGCTGCTGCTGCTCGGTCGGGACCG 12780
L A R M S F A V L P L A L L L S V R D A

12781 ACGGGGTCTGTGCGCTGCGCGGACTGACCTCCGGCGCGCTGTCGCGCAGCTCAGCTG 12840
T G S F A V A G L T S G A L S A T L T L

12841 TTGCGCGCCCGCGCGCGCGCTGATCGACCGCGGGGCTACGGTCCGGACTGGTCCGG 12900
F A P A R A R L I D R R G S R S G L V R

12901 CTGACCGTCCCGTACCTCTGCGGGCTCGCGCTGCTGATCACATTGCGCGAGGCGAAGG 12960
L T V P Y L L G L A V L I T L A E A E A

12961 CCCACCGCGCGCTGCTGCTCGCGCGCGCGGTGCGCGCGCTGTTGCGCGCGCGCTCGGT 13020
P T A A L L V A A A V A G V F A P P L G

13021 CGGACCATGCGGTGCTGTGGGCGAGGATCTGACCGCGCTGACGCCCTGCTGCACACC 13080
P T M R V L W A R I L H G R Q P L L H T

13081 GCTACGCGCTCGACTCCGTACCGGAGGAGTGTCTTACCGTGGGCGCGTGTGCGG 13140
A Y A L D S V T E E V V F T V G P L L A

13141 GCGGCGTGTGCGGTGCGGCGACCGCTCGCGTGTGATGATCAGGTGATGCTGTGATC 13200
G G L I A V A A P L A S M I T V M V L I

13201 GCGGCGGTACCGCTGCTCTGCTGTGCGCGCGGACCGCGCGCGCGCGCTGCGGC 13260
A A G T A C F V L S A A T A A A P A S G

13261 GAAGCCGACGAGGACCGCGCGCACGCGCGGCCATGCTGCGCGCGGATGCGCACGATC 13320
E A D E D R P H G R F M A L P G M R T I

13321 GTGCTGCTCTCGCGCGCTCGGCGCTGCTGCTGCGCGCGTCCAGGTGCTGCTGCGGTG 13380
T L S F G G V G L V V G V L Q V V L P F

13381 ATCGCGACCAAGCGGCTGCGCGCGCGCGCGGCGGATCTGCTGTCCATGCTGTGCGG 13440
I A D H A G S P G A G G I L L S M L S A

13441 GGCAGCGGCTCGCGCGCTCGGCTACGGCGGATGCGCTGCGCGCTCGACCGCGTGGG 13500
G S A V G G L A Y G R I A W R S T F V R

13501 CGGTGCTGCTGCTCTCACCGGTTACGCTGCGCGGTGCTGCGCTGCTGCTGACCGG 13560
R F V V L V T G F T L A V L P L C L T A

13561 AGCCGGTGGCGCGGGGCTTGGCCCTCTCGTGGGACTGCGCTGCGCGCTGCTGTC 13620
S P V P A G A F A L L V G L C L A P L F

13621 ACCACCGCTACCTGCTGCTACGAGCTGTTGAGCGCGTGGGACCGACACCGAG 13680
T T A Y L L V N D L V T A S G T A P T E

13681 GCCAACACCTGGGTCTCCACGGCCCAATAACGGAGGGTTGCGCGCGGAGCGCGCGCGCC 13740
A N T W V S T A N N G G F A A G S A A A

13741 GGTGTGCTGTCTGACTCCCGGGGCCACCGTCAACGTCACCGCGCGGTTCGCGGTGCGC 13800
G V L L D S R G P T V T V T A A F A V A

13801 GCGCGACCGCGCTCATGACCGTCTGCGCGCGCGGACCGTCTCTCGGCGCGGACAC 13860
A A T A V M T V L R R R T L L L G A G H

13861 CCCGAACCGCGCGGCCACACCGCGCGACCGCAACCGCGCGAAGCGAGGAGTGA 13920
P E P A A A T F A D R T A F A E A E *

13921 ACGATGTTGCCAAGAACCGCGGCACTGGTTCGCGCATCGCAACAGGGACGCCCCGG 13980
M S K N A A H W S R I R T G D A P G
(orf37)

13981 CGTGTACTCGCGGTGACTTCTACGGAAACGGCGCGCGCAAGGAACCACTTCCGCCACT 14040
V V L A V D F Y G T G R Q E A T F R H L

14041 GTGCGACCTGCTCAACGATCCGTCGAGTCTGGCAGCGGTCCCGCCCGCGCGGACG 14100
C D L L T D P V E V W H A V P F A P D G

14101 CGACTGGTCCAACGGCCACCGCGCGCGTCACTCGCGTGGTGGACGAGGGGCTGACAC 14160
D W S T A T G A G H L R W W T E G L D T

14161 GGTCTCGCGGACCGCGGTGCGGGCCCTGTCGCTACTCGCGCGCGCGGTCTTGGC 14220
V L A G R P V R A L V G Y C A G G V F A

14221 CTCGGCCCTCGCGACGCGCTGTCGAAACGGGAGGGCCACCGCGCGCGGTGCTGCTT 14280
S A L A D A L V E R E G H R P R V V L F

14281 CAACCCACGCGCGCGCGGTGCGCACGCTCACCGCGCACTTCGCGGTCTGATCGCGG 14340
N P S A P G V A T L T R D F R G L I A G

14341 CATGGACTCTCTCAACGACGGGAACGCGCGCTCTGTCGGCGGACGACCGCGATCG 14400
M D L L T D G E R A A L L A E T T A I R

14401 GCGGGCACACGCGCGCGCTGCTACGCTGCGCGAACGCTACGCGCGCGCTGTACCG 14460
R A H A P D A L V P V A E R Y A A L Y R

14461 CGAGGGCTGCGACCTCTGTGCGAGCGGCTCGCGTGGACGCTCTCTCGCGCGCGAAT 14520
E G C D L L C E R L G V D A S F G A E L

14521 GGCGCGCTCTCTCACTCTCACTGGCTACCTCACGGCGCGCTGACGCTGCCCGCAC 14580
A A V L H S Y L A Y L T A A L D V P P T

14581 CCGCTGTGGCGCGCGCGCTCTGCTCACTCCCGCGAGCACCGGACCGACGCTTAC 14640
P L W R G A V S L T S R E H Q G T D F T

14641 CGAAGTGGACACGGCTTGCACGTGCGCGGTCGCGAACTGCTGAGCTCCCCCAGGTGT 14700
D V E H G F D V A R A E L L S S P Q V V

14701 CGCGGCGCTGACCGCGCTCTCCGGAACACGAGGCGAGCGGATGACCTTACCTCGCG 14760
A A L T A L L R E H E A S R *
M T L T L R
(orf38)

14761 GACGCTTCTCTGACAGGCGCGCGGACCCCGACCGCGCGCTGCTGACGCGCGAC 14820
D A F L D Q A A R T P D A H A V V H G D

14821 ACTGTATGAAGTACCGGAACCTGGAACCTGCGGGCGCGCGCATGCGCGCGACGCTGCG 14880
T V W T Y R E L E L R A G R M A R T L A

14881 GCACGCGGCGGGCCCCGCGCACGCTGTTGGGGTACGCTTCGCGCGGCTCCGAACCG 14940
A R G A G P G T L V A V R L P R G P E P

14941 GTCGCGCGCTCTCTCGGGTCTGCTGACGGGAGCGGGCTACGTGCGCTGCGCGACGAC 15000
V A A L L A V V L T G A G Y V P L A D D

15001 GACCCCGCGACCGGTGCCGCGACATCTCGACACTGCGCGCGCGCTGCTGCTGCGC 15060
D P P D R C R H I L D D C A A A L L L A

15061 GAGCACCCCTCGCGGGACGACGACCCCTCACCCCGAGCGAGGCGCTGCGACCGCGCGC 15120
E H P S R D G R T L T P D E A L A P A R

15121 CCGTTCGACGCGGCCCCGGTGGCGGCGGACCGCGTACGTGATCTACACCTCGCGC 15180
P F D A A P V R A G D P A Y V I Y T S G

15181 TCCAGTGGCGCTCCGAAGGCGTGTGTGTGAAAGGGCGCGCTGCGCGCTACCTGGCA 15240
S S G R P K G V L V E Q G A L G A Y L A

15241 CAGGCCCCGCGCGCTACGACGGGCTGTCCGAGACGACGCTGCTGCTGCTGCTGCTC 15300
Q A R A R Y D G L S G R T V L H S S L S

15301 TTCGACATGGCCGTGACCCAGTCTGTGGGGCCGCTGCTGACGCGCGCGAGTCCACGTV 15360
F D M A V T S L W G P L V S G G A I H V

15361 CTCGACCTGAAGGCGATCGCTTCGCGCACCCAGCGCGCGCGCGCGCTCGGACGCTCG 15420
L D L K A I A S G T Q P P P A A S A R P

15421 TCCTTCTCAAGGTCACTCCGCTCCACCTGCGCGTGTGGGGCTGCTGCGGACCTCGC 15480
S F L K V T P S H L P L L G L L P D S C

15481 CTGCCCCACGGGCAACTCGTGTGCGCGGAGCGCGCTGACCGGCTCCGCGCTCGGACCC 15540
L P T G Q L V I G G E A L T G S A L G P

15541 TGGCGCGCGCGCACCCGACGTCACGCTGTCACGAGTACGGGCGCACCGGAGCGAC 15600
W R A A H P D V T V V N E Y G F T E A T

15601 GTGGCTGCTGCGCGTACACCGTCCGCGCGGTGACGCGCTGGACCGGCTGCGTCCCC 15660
V G C C A Y T V R P G D A V D P G A V P

15661 ATCGGACGCGCGTTCGCGGCGACCCGCTGTACGTGCTGCGACGCGGACGCGGCGCG 15720
I G R P F A G T R L Y V L D A D G E P V

15721 GCGGTGGCGGTGTGGGTGAAGTCTGACATCGCGGCGACAGTTGGCGCGGATACCTG 15780
A V G G V G E L H I A G D Q L A R G Y L

15781 GGGCGCCCGCGGTGACCGAGGAACGCTTGTCCCGGACCGTTGCGCGCGAGCGCTCC 15840
G R P R L T E E R F V P D P F A A D G S

15841 CGGATGTACCGCACCGGCGACCTGGTGCGGAAAGCGCGGACGCGGACCTGAGTACCTC 15900
R M Y R T G D L V R E R P D G D L E Y L

15901 GGGCGCGCGGACGGGAGGTGAAGTCTCGGGTACCGGATCGAGCGCGGCGGATCGAG 15960
G R A D G Q V K V S G Y R I E P G E I E

15961 GCGGTGCTCGCGCGCACGCGGGGTGAGGAGTGCAGCGCTGCTGCGCGTGGCGGAGCG 16020
A V L R G H A G V R D C A V V A V G E A

16021 GACGCGCGCGCGCTGCTGCGCTACGTGTATCGGACCGGACTCCCGCGCGGACCGCC 16080
D A R R L V A Y V V P D P D S P P G T A

16081 GCGCGGCGCGGACGCGGCGAGGCGCTGCGCGCTACATGTGCGCGGACGTTGCTC 16140

A P A R H A A E A L P P Y M V P A T F V

16141 ACCGTGCCCAGTGCCTCACCCTCAACGGGAAGCTCGACCGGACGCGTGCCTCCG 16200
T V P E L P L T P N G K L D R D A L P G

16201 CCCCTGCCGCGACGCCGGCCGGCGACCGCACCCCGCGAGACCTGCTGTGCGAG 16260
P P A G D A G P G D R T P A E T L L C E

16261 CTGCTGGCAAGGCCCTGGGATCCCGGAGATCGACGCCGACCGGACTTCCTGACGTCC 16320
L L A R A L G I P E I D A D A D F L T S

16321 GCGCGACCAAGCATACCGCGCTGAAGCTGCTCGCCGCGCCCGCGGTGCGCATCCG 16380
G G T S I T A L K L V A G A R R V G I R

16381 CTGGAATCAACACCGTCTGCGGGAAGCGACGCTGCGCCGATCTCGCGCCCGAGCC 16440
L E L T T V L R E R T V R R I L A A Q P

16441 GACGCGCTCGCCCTCGCGGAAGAGTGCCCGAGTGACCGGTTCCGTAACTGCTACCC 16500
D A A S P L A E G V P E * (orf39)

16501 CCCTCGCGGGATCATCCCCAGGCCCGCGCGAGGGCTCACCAACCGCGCGAGTACG 16560
L G G I I P R P R G E G L T T G A E Y D

16561 ACCTGGGCGGCTCGCGGACCGGGCCCGACTGCGTGGCGGCCACCGCGCGGACTGC 16620
L G P L G D A G P D W V R A H G P R L R

16621 GCGAGCGCTCGCCACGACGGGTGATCTGCTGACGGTCTGCCACCGACGAGAGCG 16680
E R L A T D G L I L L H G L P T D G D G

16681 GCGTCGACGGCTTCCACGAGCTGCTGCGCTCGCGCGCGACCGCGTGCCTACACCG 16740
V D G F H D V V G S V G G D P L P Y T E

16741 AGCGCTCCACCCCGCGAGCGTGTCAAGGGCAACATCTACCTCGACCGAGTACCGG 16800
R S T P R S V V K G N I Y T S T E Y P A

16801 CCGACCAAGCCATCCGATGCAACAGGAATCTCTACCGCCCGCATGCGCGTCCACCG 16860
D Q P I P M H N E N S Y A A H W P S T L

16861 TCTACTTCTCTGCCACCGCGCGGACACCGCGGGGCGACGCGATCGCGACGGCC 16920
Y F F C H T A P D T G G A T P I A D G R

16921 GCGCGTCTCTGACCTCATCCCGCGGAGTCCAGCGCGGCTTCTCCAAAGGGTCTGCT 16980
A V L D L I P A E V R R R F S Q G V V Y

16981 ACACCGTACGTTCCGCGCGACATGGGACTGAGCTGGCAGGAAGCGTTCAGACCGAGG 17040
T R T F R A D M G L S W Q E A F Q T E D

17041 ACCGCGCGACGTGGAACGCATTGCGCGCCCGCGACCGCGGAGTCTCTCGGACGGCG 17100
R G D V E R H C R A H G Q E F S W D G D

17101 AGTCTCTGCGACCCGCGACCGCGCGCGCGGCGACCGCGTTCGACCCCGCGACCGAGCG 17160
V L R T R H H R P A T A V D P G T G A E

17161 AGGTGTGGTTCAACAGGGCGCACCTGTTCACCCGTCGAGCTGGATCCGACCTGCGCC 17220
V W F N Q A H L F H P S S L D P D L R Q

17221 AGGTGCTCTGAGAGTACGGCGAGAACGGCTGCCCGCGACGCGCTGTCTCGCGAGG 17280
V L L E T Y G E N G L P R D A L F A D G

17281 GCACCCCGATCCCGACGCCGACCTGGCAACGCTCCGCGCGGCTTACACCCCGCGCGCG 17340
T P I F D A D L A T V R A A Y T R A A L

17341 TCGCGCTGCGGTGGCGAGAGGGCGACATCATGCTGCTCGACAACTGAGGATGGCCACG 17400
A L P W R E G D I M L V D N L R M A H G

17401 GCCGCGAGCCCTTACCGGCGAGCGCGCTACTCGTGCATGACCTCGGCGGACTCAT 17460
R E P P F T G E R R V L V A M T S A D S *

17461 GAGCGTGCAGCGCATCGGCAAGCGCTCCCGTCGGGGCGCTACCATCGCGCTGTC 17520

17521 TCGGCCATCACCCACCCGCGGCGAGGCAACGCGCGCTGCACTCCCGCGCGTGGTGGC 17580

17581 ACGGCAAGCGGATCACCGCGCCATGACCGCCAGCCCGTGTGTACATCTCGCGAGGCG 17640

17641 CCGCGATGACAGAGGTCCGAGGTGAACCTGATCCGCGCGCTGCGGGTGTGCTGGAGGCGC 17700
M T E V R G E L I R A L P G V L E A R
(orf40)

17701 GTGCGCGCGGCGGGGCAACGACCGCTTCTCGAAGCACGACGCTGTGTGTACGTACC 17760
A A R A G H T T A F L D A R R C V T Y R

17761 GGGAGTTGAGGCGCGCACCCGCGCGCTGCGCGGTCACTGCTGCGGTGGGGTGGCGC 17820
E L E A R T R R L A G S F G A V G G A Q

17821 AGGGGCGAGCGGGTGGCGCTCGTCAATGGGCAACCGGGGTGGAGATCGCGAGGTTTC 17880
G Q T G W R S S M G N R G G D G G G F F P

17881 CTCCCGGTGCTCGGGCGGAGCGGTAGGGGTGCGCTCGATTCCGGGGCCACGAGCGC 17940
P R C C G P E R * G C R S I P G P R T R

17941 GGAGCTCGCTACTTCTCTGACGACTGTGAGCGGTGGCGTGGTCAACGAGGAGACGCT 18000
S S R T S S T T V E R W R W S P R R R C

18001 GCTCGCGGGTCTCGCGATCGGCGGCGTACGATCTGCTGGTGGGGGTTCGAGCGCGT 18060
C R G S R D R R A Y G S W W G V R T P S

18061 CCGGAGGGAGCGGCTGCGGCGATCCACTCTTCTGAGCGGCTCGCGCGTTCGATCGGG 18120
R R E R L P A S T P S S G S R R R I R G

18121 GTGCGGCCACGAGACGACTCGGCTCGACGAGCGGCTGGATCTCTACAGCTCGG 18180
A R H G T T S A S T S R P G S S T R R G

18181 GACCAAGGGCGGAGCAAGGGCGTGGTCTGCGGCGAGCGCGCGCGCTGTGTGCTGGTGGC 18240
P R A G A R A W S A A S A P R C G P W R

18241 GCGCGCTACGTGCGCTGTGGGGTCTGGGGCGCAGGACCGGCTGTGTGCGCGCTGCC 18300
R R T C R R G V W G R R T G C C G R C P

18301 CATGTTCCAGCCTACGCGCACTCGCTGTGCTGCTCGGGTGGTGGCGCTGGGCGGAG 18360
C S T P T R T R C A C S G W W P W A R A

18361 CGGTACCTCTCGAACCGGGGCGGAGCGTCTGCGGGCGCTTGAAGAAACGCGTGCAG 18420
R T S S T G A R A S S G R L R N S G A A

18421 CGTGTGCGCGGTGTACCGGCGCACTACCGCTGCTCAGGAGCGCTTCGCGGACGCCCC 18480
S W P V Y P P P T A C S R A P S A T P P

18481 CCGGCCACCGCGGCGCTGCGACTGTGCGTCAACGGGGCTGCGCGCTGCGCGCGGCGC 18540
G H R P A C D C A S P G A A P C P F P G L

18541 TCGCGCGGACGTTGAGGAGCTGTGCGGCTCCCGCTGCTCGACGTTACGCGCATGACG 18600
R A D V E E L L G V P L L D G Y G S T E

18601 AGACCTGCGCAAGATCACCGTTGAGCGCTCGGCGCTCCCGGAGGGCGGTTGCGGG 18660
T C G K I T V E R L G G S R E G G C R

SEQ ID NO: 3 BLM gene PPTase ORFS 41

1 GGATCTCGCTACCGGACTTCGCCAGTGTGTCGCGCACCGAGCTCACGCGGACTGACGCTTCGGGCGCGC 80
81 GCCGCGCTTACGGGCTATGCAATCCCCGCGTGACCGGTACGACGGCGTCGCTTCGAGGAGGTGTGTCGTCGCTA 160
161 CCCGCGGATGTGCGCGCCCGCGCCCGCGAGCCGCTCCGGCAGATCTGCCCCAGCCCTCGACGAGCGCGGAGCC 240
241 TCTGGTGATCGCGCCCTCTGCGCTCGCGCGCGTACCGAAGCGCTTACCGAGCGCCCGGAGCGACCGCGTGAGCC 320
1 M I A A L L P S W A V T E H A F T D A P D D P V S L 26
321 TCCTCTTCCCGAGGAGCGCGCCACGTGCGCCGCGCGCTCCCAAGCGCTGACAGATGTCGCCACCCTCGGGGTGTC 400
27 L F P E E A A H V A R A V P K R L H E F A T V R V C 52
401 GCCGCGCGCCCTCGCGCGCTGCGCTCCCGCGCGCTGCTGCTGCGCGCGGACGCGCGCGCGCGCGCGCGCGCG 480
53 A R A A L G R L G L P P G P L L P G R R G A P S N P D 79
481 CGGGTGGTGGGAGCATGACGACTGTGAGGGCTTCGGGGGCGCGCGGTGCGCGCGCGCGCGCGCGCGCGCGCG 560
80 G V V G S M T H C Q G F R G A A V A R A A D A A S L G 106
561 GGATAGACGCGCGAAGCGCGCGCTCCCGGAGCGCGCTCTCGCCATGGTCTCGCTGCGCGCGCGCGCGCGCGCG 640
107 I D A E P N G P L P D G V L A M V S L P S E R E W L 132
641 GCGGAGTGGGCGCGCGCGCGCGCGGACGTGCACTGGGACCGGCTGCTGTTGAGCGCAAGGAGAGCGCTTCAAGGCGTG 720
133 A G L A A R R P D V H W D R L L F S A K E S V F K A W 159
721 GTACCCGCTGACCGGCTGGAGCTTGGACTTGCAGAGGCGGAGCTGGCGCTCGATCCGGACCGCGGAGCTTCAAGCGCC 800
160 Y P L T G L E L D F D E A E L A V D P D A G T F T A R 186
801 GGCTGCTGGTCCGGGACCGGTGCTGCGCGCGCTCGGCTGGACGGGTTCGAGGGGCGCTGGGCGCGCGGCGGGCGCT 880
187 L L V P G P V V G G R R L D G F E G R W A A E G L 212
881 GTCTGACGGCCATCGCGCTCGCGCGCGCGCGCGTACCGCGGAGGAATCGGCGGAGGGCGCGGGAAGGAGCGACTGC 960
213 V V T A I A V A A P A G T A E E S A E G A G K E A T A 239
961 GGACGACGGAACCGCGCTCCGCTAAACG 1040
240 D D R T A V P * 247
1041 GCGCGCGCGCGCGCGCGCTCCCGCGTGCAGGACGAGGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCG 1120
1121 AGTCGCGACGACGACGATTCGCTGTGGTTCGAGTTGAGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCG 1200
1201 ATGTGGACGGGATCTGGATGACGTTGCGCGAGACGACCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCGCG 1280
1281 GAGGGCGTGCCTGCGAGACG 1360
1361 GGTGACACCTTCGTTGCTGACAGGCTCGAGCTCAAGCGCTTGAAGCGCGCGCGCGCGCGCGCGCGCGCGCG 1440
1441 CGAAGGTTTCGAATCTGTCG 1520
1521 CGTGCCTATCTGACG 1600
1601 GGGAGGGCATGAGCGGATCGGATCTGCG 1680
1681 GAGCGGCTCTTCCGCTCAGAGGTGCTGCTGCTGCTCCGACGGGACCGCGGACGAGATCGCGCGCGCGCGCGCG 1760
1761 C 1761